Appendix

A1 Probability Theory

The modern formulation of probability theory is due to Kolmogorov [1933]. In that 60-page monograph, Kolmogorov introduced the notion of probability spaces, the axiomatic definition of probability, the modern definition of random variables, and more. For an excellent review of Kolmogorov's fundamental contribution, see Nualart [2004]. In this Appendix, we review concepts of probability theory at the graduate level, including many concepts that are needed in the book. The language of measure theory is used, although measure-theoretical concepts are only needed in the book in the starred additional topics sections. For excellent book-length treatments of probability theory, the reader is referred to Billingsley [1995], Chung [1974], Loève [1977], Cramér [1999], and Rosenthal [2006], while a thorough elementary non-measure-theoretical introduction is provided by Ross [1994].

A1.1 Sample Space and Events

A *sample space* S is the set of all outcomes of an experiment. A *σ-algebra* is a collection $\mathcal{F}$ of subsets of S that is closed under complementation, (countable) intersection, and (countable) union. Each set E in $\mathcal{F}$ is called an *event*. Hence, complementation of events are events, and (countable) unions and intersections of events are events.

Event E is said to *occur* if it contains the outcome of the experiment. Whenever $E \subseteq F$ for two events E and F, the occurrence of E implies the occurrence of F. The complement event E^c is an event, which occurs iff(if and only if) E does not occur. The union $E \cup F$ is an event, which occurs iff E, F, or both E and F occur. On the other hand, the intersection $E \cap F$ is also an event, which occurs iffboth E and F occur. Finally, if $E \cap F = \emptyset$ (the latter is called the *impossible event*), then E or F may occur but not both.

U. Braga-Neto, *Fundamentals of Pattern Recognition and Machine Learning*,
https://doi.org/10.1007/978-3-030-27656-0

For example, if the experiment consists of flipping two coins, then

$$S = \{(H,H),\,(H,T),\,(T,H),\,(T,T)\}\,. \tag{A.1}$$

In this case, the σ-algebra contains all subsets of S (any subset of S is an event); e.g., event E that the first coin lands tails is: $E = \{(T,H),\,(T,T)\}$. Its complement E^c is the event that the first coin lands heads: $E^c = \{(H,H),\,(H,T)\}$. The union of these two events is the entire sample space S: one or the other must occur. The intersection is the impossible event: the coin cannot land both heads and tails on the first flip.

If on the other hand, the experiment consists in measuring the lifetime of a lightbulb, then

$$S = \{t \in R \mid t \geq 0\}\,. \tag{A.2}$$

Here, for reasons that will be described later, it is not desirable to consider all possible subsets in S as events. Instead, we consider the smallest σ-algebra that contains all intervals in S; this is called the *Borel σ-algebra* in S, and the events in it are called *Borel sets*; e.g., the event that the lightbulb will fail at or earlier than t time units is the Borel set $E = [0,t]$. The entire sample space is the countable union $\bigcup_{t=1}^{\infty} E_t$, where $\{E_t; t = 1,2,\ldots\}$ is called an increasing *sequence of events*. Borel sets can be quite complicated (e.g., the famous Cantor set is a Borel set). There are sets of real numbers that are not Borel sets, but these are quite exotic and of no real interest. Generalizing, the Borel σ-algebra $\mathcal{B}^d$ of R^d is the smallest σ-algebra of subsets of R^d that contains all rectangular volumes in R^d. If $d = 1$, we write $\mathcal{B}^1 = \mathcal{B}$.

Limiting events are defined as follows. Given any sequence $\{E_n; n = 1,2,\ldots\}$ of events, the *lim sup* is defined as:

$$\limsup_{n\to\infty} E_n \;=\; \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i\,. \tag{A.3}$$

We can see that $\limsup_{n\to\infty} E_n$ occurs iff E_n occurs for an infinite number of n, that is, E_n *occurs infinitely often*. This event is also denoted by $[E_n\, i.o.]$. On the other hand, the *lim inf* is defined as:

$$\liminf_{n\to\infty} E_n \;=\; \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} E_i\,. \tag{A.4}$$

We can see that $\liminf_{n\to\infty} E_n$ occurs iff E_n occurs for all but a finite number of n, that is, E_n *eventually occurs for all* n. Clearly, $\liminf_{n\to\infty} E_n \subseteq \limsup_{n\to\infty} E_n$. If the two limiting events coincide, then we define

$$\lim_{n\to\infty} E_n \;=\; \liminf_{n\to\infty} E_n \;=\; \limsup_{n\to\infty} E_n\,. \tag{A.5}$$

Notice that, if $E_1 \subseteq E_2 \subseteq \ldots$ (an increasing sequence), then

$$\lim_{n\to\infty} E_n \;=\; \bigcup_{n=1}^{\infty} E_n\,, \tag{A.6}$$

whereas, if $E_1 \supseteq E_2 \supseteq \ldots$ (a decreasing sequence), then

$$\lim_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} E_n \,. \tag{A.7}$$

A *measurable space* $(S, \mathcal{F})$ is a pair consisting of a set S and a σ-algebra defined on it. For example, $(R^d, \mathcal{B}^d)$ is the standard *Borel-measurable space*. A *measurable* function between two measurable spaces $(S, \mathcal{F})$ and $(T, \mathcal{G})$ is defined to be a mapping $f : S \to T$ such that for every $E \in \mathcal{G}$, the pre-image

$$f^{-1}(E) = \{x \in S \mid f(x) \in E\} \tag{A.8}$$

belongs to $\mathcal{F}$. A function $f : R^d \to R^k$ is said to be *Borel-measurable* if it is a measurable function between $(R^d, \mathcal{B}^d)$ and $(R^k, \mathcal{B}^k)$. A Borel-measurable function is a very general function. For our purposes, it can be considered to be an arbitrary function. In this book, all functions (including classifiers and regressions) are assumed to be Borel-measurable.

A1.2 Probability Measure

A *measure* on $(S, \mathcal{F})$ is a real-valued function μ defined on each $E \in \mathcal{F}$ such that

A1. $0 \le \mu(E) \le \infty$,

A2. $\mu(\emptyset) = 0$,

A3. Given any sequence $\{E_n; n = 1, 2, \ldots\}$ in $\mathcal{F}$ such that $E_i \cap E_j = \emptyset$ for all $i \neq j$,

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i) \quad (\sigma\text{-additivity}) \,. \tag{A.9}$$

The triple $(S, \mathcal{F}, \mu)$ is called a *measure space*. A *probability measure* P is a measure such that $P(S) = 1$. A *probability space* is a triple $(S, \mathcal{F}, P)$, consisting of a sample space S, a σ-algebra $\mathcal{F}$ containing all the events of interest, and a probability measure P. A probability space is a model for a stochastic experiment; the properties of the latter are completely determined once a probability space is specified.

Lebesgue measure on $(R^d, \mathcal{B}^d)$ is a measure λ that agrees with the usual definition of length of intervals in R, $\lambda([a, b]) = b - a$, area of rectangles in R^2, $\lambda([a, b] \times [c, d]) = (b - a)(d - c)$, and so on for higher-dimensional spaces, and uniquely *extends* it to complicated (Borel) sets. Notice that $\lambda(\{x\}) = 0$, for all $\mathbf{x} \in R^d$, since a point has no spatial extension (it follows that it makes no difference whether intervals and rectangles are open, closed, or half-open). By σ-additivity,

any countable subset of R^d has Lebesgue measure zero, and there are uncountable sets that have Lebesgue measure zero as well (e.g., the Cantor set in R). Sets of Lebesgue measure zero are very sparse; any property that holds in R^d outside of such a set is said to hold *almost everywhere* (a.e.). The measure space $(R^d, \mathcal{B}^d, \lambda)$ provides the standard setting for mathematical analysis.

Lebesgue measure restricted to $([0,1], \mathcal{B}_0)$, where $\mathcal{B}_0$ is the σ-algebra containing all Borel subsets of $[0,1]$, is a probability measure, since $\lambda([0,1]) = 1$. The probability space $([0,1], \mathcal{B}_0.\lambda)$ provides a model for the familiar uniform distribution on $[0,1]$. A famous impossibility theorem states that there does not exist a probability measure defined on $([0,1], 2^{[0,1]})$, where $2^{[0,1]}$ denotes the σ-algebra of all subsets of $[0,1]$, such that $P(\{x\}) = 0$ for all $x \in [0,1]$ [Billingsley, 1995, p. 46]. Therefore, λ cannot be extended to all subsets of $[0,1]$. This shows the need to restrict attention to the σ-algebra of Borel sets, where a unique extension of λ exists. (Lebesgue measure can be uniquely extended to even more general sets, but this is not of interest here.)

The following properties of a probability measure are straightforward consequences of axioms A1–A3 plus the requirement that $P(S) = 1$:

P1. $P(E^c) = 1 - P(E)$.

P2. If $E \subseteq F$ then $P(E) \leq P(F)$.

P3. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

P4. (Union Bound) For any sequence of events $E_1, E_2, \ldots$

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} P(E_n)\,. \tag{A.10}$$

P5. (Continuity from below.) If $\{E_n; n = 1, 2, \ldots\}$ is an increasing sequence of events, then

$$P(E_n) \uparrow P\left(\bigcup_{n=1}^{\infty} E_n\right) \tag{A.11}$$

P6. (Continuity from above.) If $\{E_n; n = 1, 2, \ldots\}$ is an decreasing sequence of events, then

$$P(E_n) \downarrow P\left(\bigcap_{n=1}^{\infty} E_n\right) \tag{A.12}$$

Using P5 and P6 above, it is easy to show that

$$P\left(\liminf_{n\to\infty} E_n\right) \leq \liminf_{n\to\infty} P(E_n) \leq \limsup_{n\to\infty} P(E_n) \leq P\left(\limsup_{n\to\infty} E_n\right). \tag{A.13}$$

From this, the general *continuity of probability measure* property follows: for any sequence of events $\{E_n; n = 1, 2, \ldots\}$,

$$P\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} P(E_n)\,. \tag{A.14}$$

In some cases, it can be easy to determine the probability of limsup and liminf events. For example, it follows from (A.13) that mere convergence of $P(E_n)$ to 1 or 0 as $n \to \infty$ implies that $P(\limsup_{n\to\infty} E_n) = 1$ and $P(\liminf_{n\to\infty} E_n) = 0$, respectively. In the general case, it may not be simple to determine the value of these probabilities. The *Borel-Cantelli Lemmas* give sufficient conditions for the probability of limsup to be 0 and 1 (through the identity $P(\liminf E_n) = 1 - P(\limsup E^c)$, corresponding results on the probability of liminf can be derived).

Theorem A.1. *(First Borel-Cantelli Lemma.) For* any *sequence of events* $E_1, E_2, \ldots$

$$\sum_{n=1}^{\infty} P(E_n) < \infty \;\Rightarrow\; P([E_n\ i.o.]) = 0\,. \tag{A.15}$$

Proof. Continuity of probability measure and the union bound allow one to write

$$P([E_n\ i.o.]) = P\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right) = P\left(\lim_{n\to\infty}\bigcup_{i=n}^{\infty} E_i\right) = \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right) \le \lim_{n\to\infty}\sum_{i=n}^{\infty} P(E_i)\,. \tag{A.16}$$

But if $\sum_{n=1}^{\infty} P(E_n) < \infty$ then the last limit must be zero, proving the claim. ⋄

The converse to the First Lemma holds if the events are independent.

Theorem A.2. *(Second Borel-Cantelli Lemma.) For an* independent *sequence of events* $E_1, E_2, \ldots,$

$$\sum_{n=1}^{\infty} P(E_n) = \infty \;\Rightarrow\; P([E_n\ i.o.]) = 1 \tag{A.17}$$

Proof. By continuity of probability measure,

$$P([E_n\ i.o.]) = P\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right) = P\left(\lim_{n\to\infty}\bigcup_{i=n}^{\infty} E_i\right) = \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right) = 1 - \lim_{n\to\infty} P\left(\bigcap_{i=n}^{\infty} E_i^c\right), \tag{A.18}$$

where the last equality follows from DeMorgan's Law. Now, by independence,

$$P\left(\bigcap_{i=n}^{\infty} E_i^c\right) = \prod_{i=n}^{\infty} P(E_i^c) = \prod_{i=n}^{\infty} (1 - P(E_i)) \tag{A.19}$$

From the inequality $1 - x \le e^{-x}$ we obtain

$$P\left(\bigcap_{i=n}^{\infty} E_i^c\right) \le \prod_{i=1}^{\infty} \exp(-P(E_i)) = \exp\left(-\sum_{i=n}^{\infty} P(E_i)\right) = 0 \tag{A.20}$$

since, by assumption, $\sum_{i=n}^{\infty} P(E_i) = \infty$, for all n. From (A.18) and (A.20), $P([E_n\ i.o.]) = 1$, as required. ⋄

A1.3 Conditional Probability and Independence

Given that an event F has occurred, for E to occur, $E \cap F$ has to occur. In addition, the sample space gets *restricted* to those outcomes in F, so a normalization factor $P(F)$ has to be introduced. Therefore, assuming that $P(F) > 0$,

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}. \tag{A.21}$$

For simplicity, it is usual to write $P(E \cap F) = P(E, F)$ to indicate the *joint probability* of E and F. From (A.21), one then obtains

$$P(E, F) = P(E \mid F)P(F), \tag{A.22}$$

which is known as the *multiplication rule*. One can also condition on multiple events:

$$P(E \mid F_1, F_2, \ldots, F_n) = \frac{P(E \cap F_1 \cap F_2 \cap \ldots \cap F_n)}{P(F_1 \cap F_2 \cap \ldots \cap F_n)}. \tag{A.23}$$

This allows one to generalize the multiplication rule thus:

$$P(E_1, E_2, \ldots, E_n) = P(E_n \mid E_1, \ldots, E_{n-1})P(E_{n-1} \mid E_1, \ldots, E_{n-2}) \cdots P(E_2 \mid E_1)P(E_1). \tag{A.24}$$

The *Law of Total Probability* is a consequence of axioms of probability and the multiplication rule:

$$P(E) = P(E, F) + P(E, F^c) = P(E \mid F)P(F) + P(E \mid F^c)(1 - P(F)). \tag{A.25}$$

This property allows one to compute a hard unconditional probability in terms of easier conditional ones. It can be extended to multiple conditioning events via

$$P(E) = \sum_{i=1}^{n} P(E, F_i) = \sum_{i=1}^{n} P(E \mid F_i)P(F_i), \tag{A.26}$$

for pairwise disjoint F_i such that $\bigcup F_i \supseteq E$.

One of the most useful results of probability theory is *Bayes Theorem*:

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)} = \frac{P(F \mid E)P(E)}{P(F \mid E)P(E) + P(F \mid E^c)(1 - P(E)))} \tag{A.27}$$

Bayes Theorem can be interpreted as a way to (1) "invert" the probability $P(F \mid E)$ to obtain the probability $P(E \mid F)$; or (2) "update" the "prior" probability $P(E)$ to obtain the "posterior" probability $P(E \mid F)$.

Events E and F are independent if the occurrence of one does not carry information as to the occurrence of the other. That is, assuming that all events have nonzero probability,

$$P(E \mid F) = P(E) \text{ and } P(F \mid E) = P(F). \tag{A.28}$$

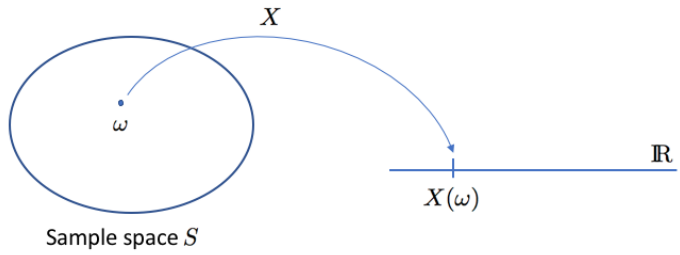

Figure A.1: A real-valued random variable.

It is easy to see that this is equivalent to the condition

$$P(E,F) \;=\; P(E)P(F)\,. \tag{A.29}$$

If E and F are independent, so are the pairs (E,F^c), (E^c,F), and (E^c,F^c). However, E being independent of F and G does not imply that E is independent of $F\cap G$. Furthermore, three events E, F, G are independent if $P(E,F,G) = P(E)P(F)P(G)$ and each pair of events is independent. This can be extended to independence of any number of events, by requiring that the joint probability factor and that all subsets of events be independent.

Finally, we remark that $P(\cdot|F)$ is a probability measure, so that it satisfies all properties mentioned previously. In particular, it is possible to define the notion of conditional independence of events.

A1.4 Random Variables

A random variable can be thought of roughly as a "random number." Formally, a random variable X defined on a probability space $(S,\mathcal{F},P)$ is a measurable function X between $(S,\mathcal{F})$ and $(R,\mathcal{B})$ (see Section A1.1 for the required definitions). Thus, a random variable X assigns to each outcome $\omega\in S$ a real number $X(\omega)$ — see Figure A.1 for an illustration.

By using properties of the inverse set function, it is easy to see that the set function

$$P_X(B) \;=\; P(X\in B) \;=\; P(X^{-1}(B))\,, \quad \text{for } B\in\mathcal{B}\,, \tag{A.30}$$

is a probability measure on $(R,\mathcal{B})$, called the *distribution* or *law* of X. (Note that P_X is well defined, since X is assumed measurable, and thus $X^{-1}(B)$ is an event in $\mathcal{F}$, for each $B\in\mathcal{B}$.) If $P_X = P_Y$ then X and Y are *identically distributed*. This does not mean they are identical: take X and Y to be uniform over $[0,1]$ with $Y = 1-X$. In this case, $P_X = P_Y$ but $P(X=Y) = 0$. On the other hand, if $P(X=Y)=1$, then X and Y are identically distributed.

An alternative characterization of a random variable X is provided by the *cumulative distribution function* (CDF) $F_X : R \to [0, 1]$, defined by

$$F_X(x) \;=\; P_X((-\infty, x]) \;=\; P(X \leq x), \quad x \in R. \tag{A.31}$$

It can be seen that the CDF has the following properties:

F1. F_X is non-decreasing: $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$.

F2. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.

F3. F_X is right-continuous: $\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$.

The following remarkable theorem states that the information in the *set function* P_X is equivalent to the information in the *point function* F_X; for a proof, see [Rosenthal, 2006, Prop. 6.0.2].

Theorem A.3. *Let X and Y be two random variables (possibly defined on two different probability spaces). Then $P_X = P_Y$ if and only if $F_X = F_Y$.*

Furthermore, it can be shown that given a probability measure P_X on $(R, \mathcal{B})$, there is a random variable X defined on some probability space that has P_X for its distribution; and equivalently, given any function F_X satisfying properties F1-F3 above, there is an X that has F_X as its CDF [Billingsley, 1995, Thm 14.1].

If $X_1, \ldots, X_n$ are *jointly-distributed* random variables (i.e., defined on the same probability space) then they are said to be independent if

$$P(\{X_1 \in B_1\} \cap \ldots \cap \{X_n \in B_n\}) \;=\; P_{X_1}(B_1) \cdots P_{X_n}(B_n)\,, \tag{A.32}$$

for any Borel sets $B_1, \ldots, B_n$. Equivalently, they are independent if

$$P(\{X_1 \leq x_1\} \cap \ldots \cap \{X_n \leq x_n\}) \;=\; F_{X_1}(x_1) \cdots F_{X_n}(x_n)\,, \tag{A.33}$$

for any points $x_1, \ldots x_n \in R$. If in addition $P_{X_1} = \cdots = P_{X_n}$, or equivalently, $F_{X_1} = \cdots = F_{X_n}$, then $X_1, \ldots, X_n$ are *independent and identically distributed* (i.i.d.) random variables.

Discrete Random Variables

If the distribution of a random variable X is concentrated on a countable number of points $x_1, x_2, \ldots$, i.e., $P_X(\{x_1, x_2, \ldots\}) = 1$, then X is said to be a *discrete* random variable. For example, let X be the numerical outcome of the roll of a six-sided. Then P_X is concentrated on the set $\{1, 2, 3, 4, 5, 6\}$.

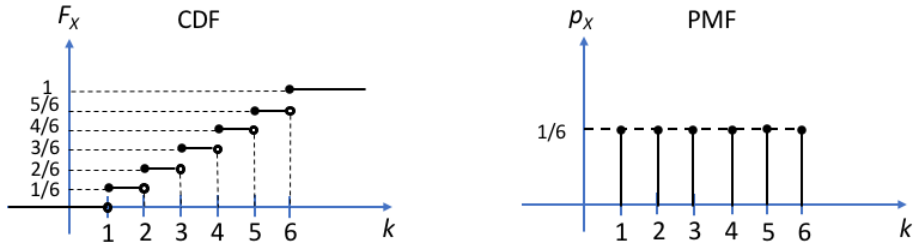

Figure A.2: The CDF and PMF of a uniform discrete random variable.

The CDF F_X for this example can be seen in Figure A.2. As seen in this plot, F_X is a "staircase" function, with "jumps" located at the points masses in P_X. This is a general fact for any discrete random variable X.

A discrete random variable X can thus be completely specified by the location and size of the jumps in F_X (since that specifies F_X). In other words, a discrete random variable X is specified by its *probability mass function* (PMF), defined by

$$p_X(x_k) \;=\; P(X = x_k) \;=\; F_X(x_k) - F_X(x_k-)\,, \tag{A.34}$$

at all points $x_k \in R$ such that $P_X(\{x_k\}) > 0$. See Figure A.2 for the PMF in the previous die-rolling example.

Clearly, discrete random variables $X_1, \ldots, X_n$ are independent if

$$P(\{X_1 = x_{k_1}\} \cap \ldots \cap \{X_n = x_{k_n}\})\,, = \; p_{X_1}(x_{k_1}) \cdots p_{X_n}(x_{k_n}) \tag{A.35}$$

at all sets of points where the corresponding PMFs are defined.

Useful discrete random variables include the already mentioned uniform r.v. over a finite set of numbers K with PMF

$$p_X(x_k) \;=\; \frac{1}{|K|}\,, \quad k \in K\,, \tag{A.36}$$

the Bernoulli with parameter $0 < p < 1$, with PMF

$$\begin{aligned} p_X(0) \;&=\; 1 - p\,, \\ p_X(1) \;&=\; p\,, \end{aligned} \tag{A.37}$$

the Binomial with parameters $n \in \{1, 2, \ldots\}$ and $0 < p < 1$, such that

$$p_X(x_k) \;=\; \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\,, \tag{A.38}$$

the Poisson with parameter $\lambda > 0$, such that

$$p_X(x_k) \;=\; e^{-\lambda}\frac{\lambda^k}{k!}\,, \quad k = 0, 1, \ldots \tag{A.39}$$

and the Geometric with parameter $0 < p < 1$ such that

$$p_X(x_k) \;=\; (1-p)^{k-1}p\,, \quad k = 1, 2, \ldots \tag{A.40}$$

A binomial r.v. with parameters n and p has the distribution of a a sum of n i.i.d. Bernoulli r.v.s with parameter p.

Continuous Random Variables

The transition from discrete to continuous random variables is nontrivial. A continuous random variable X should have the following two smoothness properties:

C1. F_X is continuous, i.e., it contains no jumps; i.e., $P(X = x) = 0$ for all $x \in R$.

C2. There is a nonnegative function p_X such that

$$P(a \le X \le b) \;=\; F_X(b) - F_X(a) \;=\; \int_a^b p_X(x)\,dx\,, \tag{A.41}$$

for $a, b \in R$, with $a \le b$. In particular, $\int_{-\infty}^{\infty} p_X(x)\,dx = 1$.

It follows from the properties of the integral that C2 implies C1. However, it is one of the surprising facts of probability theory that C1 does not imply C2: there are continuous CDFs that do not satisfy C2. The counterexamples are admittedly exotic. For instance, the *Cantor function* is a continuous increasing function defined on the interval $[0, 1]$, which has derivative equal to zero on the complement of the Cantor set, i.e., almost everywhere, but grows continuously from 0 to 1. The Cantor function is constant almost everywhere, but manages to grow continuously, without jumps. Such functions are called *singular* (or "devil staircases" in the popular literature). The Cantor function (suitably extended outside the interval $[0, 1]$) defines a continuous CDF that cannot satisfy C2. Such exotic examples can be ruled out if one requires the CDF to have a smoothness property known as *absolute continuity* (which is more stringent than simple continuity). In fact, it can be shown that absolute continuity of F_X is *equivalent* to C2. It is also equivalent to the requirement that $P(X \in B) = 0$ for any Borel set B of measure zero, not simply on isolated points, as in C1, or countable set of points. It can indeed be shown that any CDF can be decomposed uniquely into a sum of a discrete, singular, and absolute continuous components.[1]

[1]For proofs and more details, the reader is referred to Sections 31 and 32 of Billingsley [1995] and Chapter 1 of Chung [1974]. The construction of the Cantor function is described in Chapter 7 of Schroeder [2009].

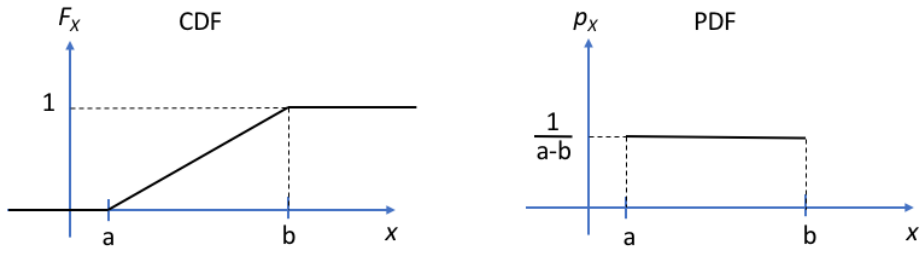

Figure A.3: The CDF and PDF of a uniform continuous random variable.

The definition of a continuous random variable X requires F_X to be absolutely continuous, not simply continuous, in which case C2 is satisfied, and p_X is called a *probability density function* (PDF). (Perhaps it would be more appropriate to call these *absolutely continuous* random variables, but the terminology "continuous random variable" is entrenched.) See Figure A.3 for an illustration of the CDF and PDF of a uniform continuous random variable. The CDF of a continuous random variable does not have to be differentiable everywhere (in this example, it fails to be differentiable at a and b). But where it is differentiable, $dF_X(x)/dx = p_X(x)$ (the density can take arbitrary values where F_X is not differentiable, and this happens at most over a set of Lebesgue measure zero).

Useful continuous random variables include the already mentioned uniform r.v. over the interval $[a, b]$, with density

$$p_X(x) = \frac{1}{b-a}, \quad a < x < b, \tag{A.42}$$

the univariate Gaussian r.v. with parameters μ and $\sigma > 0$, such that

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in R, \tag{A.43}$$

the exponential r.v. with parameter $\lambda > 0$, such that

$$p_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0 \tag{A.44}$$

the gamma r.v. with parameters $\lambda, t > 0$, such that

$$p_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{t-1}}{\Gamma(t)}, \quad x \geq 0, \tag{A.45}$$

where$\Gamma(\ t) = \int_0^\infty e^{-u}u^{t-1}du$, and the beta r.v. with parameters $a, b > 0$, such that:

$$p_X(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1, \tag{A.46}$$

where $B(a, b) = \Gamma(\ a + b)/\Gamma(a)\Gamma(b)$. Among these, the Gaussian is the only one defined over the entire real line; the exponential and gamma are defined over the nonnegative real numbers, while the uniform and beta have bounded support. In fact, the uniform r.v. over $[0, 1]$ is just a beta with $a = b = 1$, while an exponential r.v. is a gamma with $t = 1$.

General Random Variables

There are random variables that are neither continuous nor discrete. Of course, an example of that is afforded by a mixture of a discrete random variable and a continuous random variable. The CDF of such a mixed random variable has jumps, but it is not a staircase function. However, there are more general random variables that are *not* mixtures of this kind; e.g., the random variable corresponding to the Cantor CDF.

A1.5 Joint and Conditional Distributions

The *joint CDF* of two jointly-distributed random variables X and Y is a function $F_{XY} : R \times R \to [0, 1]$ defined by

$$F_{XY}(x, y) \;=\; P(\{X \le x\} \cap \{Y \le y\}) \;=\; P(X \le x, Y \le y), \quad x, y \in R\,. \tag{A.47}$$

This is the probability of the "lower-left quadrant" with corner at (x, y). Note that $F_{XY}(x, \infty) = F_X(x)$ and $F_{XY}(\infty, y) = F_Y(y)$. These are called the *marginal CDFs*.

If X and Y are jointly-distributed continuous random variables, then we define the *joint density*

$$p_{XY}(x, y) \;=\; \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad x, y \in R\,, \tag{A.48}$$

at all points where the derivative is defined. The joint density function $p_{XY}(x, y)$ integrates to 1 over R^2. The *marginal densities* are given by

$$\begin{aligned} p_X(x) \;&=\; \int_{-\infty}^{\infty} p_{XY}(x, y)\, dy\,, \quad x \in R\,, \\ p_Y(y) \;&=\; \int_{-\infty}^{\infty} p_{XY}(x, y)\, dx\,, \quad y \in R\,, \end{aligned} \tag{A.49}$$

The random variables X and Y are *independent* if $p_{XY}(x, y) = p_X(x)p_Y(y)$, for all $x, y \in R$. It can be shown that if X and Y are independent and $Z = X + Y$ then

$$p_Z(z) \;=\; \int_{-\infty}^{\infty} p_X(x) p_Y(z - x)\, dx\,, \quad z \in R\,, \tag{A.50}$$

with a similar expression in the discrete case for the corresponding PMFs. The above integral is known as the *convolution integral*.

If $p_Y(y) > 0$, the *conditional density* of X given $Y = y$ is defined by:

$$p_{X|Y}(x \mid y) \;=\; \frac{p_{XY}(x, y)}{p_Y(y)}, \quad x \in R\,. \tag{A.51}$$

For an event E, the conditional probability $P(E \mid Y = y)$ needs care if Y is a continuous random variable, as $P(Y = y) = 0$. But as long as $p_Y(y) > 0$, this probability can be defined (the details are outside of the scope of this review):

$$P(E \mid Y = y) \;=\; \int_E p_{X|Y}(x \mid y)\, dx\,. \tag{A.52}$$

The "Law of Total Probability" for random variables is a generalization of (A.26):

$$P(E) \;=\; \int_{-\infty}^{\infty} P(E \mid Y = y)\, p_Y(y)\, dy\,. \tag{A.53}$$

The concepts of joint PMF, marginal PMFs, and conditional PMF can defined in a similar way. For conciseness, this is omitted in this review.

A1.6 Expectation

The expectation of a random variable has several important interpretations: 1) its average value (weighted by the probabilities); 2) a summary of its distribution (sometimes referred to as a "location parameter"); 3) a prediction of its future value. The latter meaning is the most important one for pattern recognition and machine learning.

Expectation can be formalized by using the notion of integration, which we briefly review next. For a measure space $(S, \mathcal{F}, \mu)$ and a Borel-measurable function $f : S \to R$, one defines the integral

$$\int f\, d\mu \;=\; \int f(\omega)\, \mu(d\omega) \tag{A.54}$$

as a number in $R \cup \{-\infty\,, \infty\}$, as follows. First, if $f = I_A$ is the indicator of a set $A \in \mathcal{F}$, then $\int f\, d\mu = \mu(A)$, i.e., integrating a constant "1" over a set produces just the measure of that set. Next, if $f = \sum_{i=1}^n x_i I_{A_i}$, where the $x_i \in R$ and the A_i are measurable sets that partition S, then

$$\int f\, d\mu \;=\; \sum_{i=1}^{n} x_i \mu(A_i)\,. \tag{A.55}$$

Such a function f is called *simple*, as it takes on a finite number of values $x_1, \ldots, x_n$, with $f^{-1}(\{x_i\}) = A_i$, for $i = 1, \ldots, n$. Next, for general nonnegative function f, one defines its integral as

$$\int f\, d\mu \;=\; \sup \left\{ \int g\, d\mu \;\mid\; g : S \to R \text{ is simple and } g \le f \right\}\,. \tag{A.56}$$

Finally, for general f, define nonnegative functions $f^+(\omega) = f(\omega) I_{f(\omega)>0}$ and $f^-(\omega) = -f(\omega) I_{f(\omega)\le 0}$. Clearly, $f = f^+ - f^-$, so the integral of f is defined as

$$\int f\, d\mu \;=\; \int f^+\, d\mu \;-\; \int f^-\, d\mu\,, \tag{A.57}$$

provided that at least one of $\int f^+ \, d\mu$ and $\int f^- \, d\mu$ is finite. If both are finite, then $-\infty < \int f \, d\mu < \infty$, and f is said to be *integrable* with respect to measure μ. Since $|f| = f^+ + f^-$, f is integrable if and only if $\int |f| \, d\mu < \infty$. If $\int f^+ \, d\mu = \int f^- \, d\mu = \infty$, then the integral of f is not defined at all.

The integral ignores everything that happens over sets of measure zero: if $f = g$ outside a set of measure zero, then $\int f \, d\mu = \int g \, d\mu$. Hence, if $f = 0$ a.e., then $\int f \, d\mu = 0$, and the integral of nonnegative f is positive if and only if $f > 0$ over a set of nonzero measure.

The integral of f over a set $A \in \mathcal{F}$ is defined as $\int_A f \, d\mu = \int I_A f \, d\mu$, if it exists. If f is nonnegative, then $\nu(A) = \int_A f \, d\mu$ defines a measure on $(S, \mathcal{F})$, and f is called a *density* of ν with respect to μ (densities are unique up to sets of μ-measure zero). It is clear that $\nu(A) = 0$ whenever $\mu(A) = 0$; any measure ν with this property is said to be *absolutely continuous* with respect to μ (this is a generalization of the previous definition, as we comment below). The following theorem can be proved by showing that it holds for indicators, simple functions, and then nonnegative functions through (A.56).

Theorem A.4. *If $g : S \to R$ is integrable and $f : S \to R$ is a density of ν with respect to μ, then*

$$\int g(\omega) \, \nu(d\omega) \; = \; \int g(\omega) f(\omega) \, \mu(d\omega) \, . \tag{A.58}$$

The general integral has all the properties with which one if familiar in Calculus, such as linearity: it can be shown that if f and g are integrable and a and b are constants, then

$$\int (af + bg) \, d\mu \; = \; a \int f \, d\mu \; + \; b \int g \, d\mu \, . \tag{A.59}$$

If the measure space is $(R, \mathcal{B}, \lambda)$ then the integral of a function $f : R \to R$,

$$\int f \, d \, \lambda = \; \int f(x) \, \lambda(dx) \tag{A.60}$$

is the *Lebesgue integral* of f, if it exists. It can be shown that the Lebesgue integral coincides with the usual Riemann integral, whenever the latter exists. But the full generality of the Lebesgue integral is needed to integrate complicated functions, or functions over complicated sets. The classical example is provided by the function $f : R \to R$ defined as $f(x) = 1$ if x is rational, and $f(x) = 0$, otherwise. Notice that $f = I_Q$, the indicator of the set of rationals Q. This function is extremely irregular (discontinuous and nondifferentiable at every point) and not Riemann-integrable. However, f is measurable and Lebesgue-integrable, with $\int f(x) \, \lambda(dx) = \lambda(Q) = 0$. All integrals mentioned before in this Appendix, including (A.41), should be considered to be Lebesgue integrals.

Now, given a random variable X defined on a probability space $(S, \mathcal{F}, P)$, the expectation $E[X]$ is simply the integral of X over S according to the probability measure P:

$$E[X] \; = \; \int X \, dP \; = \; \int X(\omega) \, P(d\omega) \, , \tag{A.61}$$

if it exists. So expectation is an integral, and all definitions and properties mentioned previously in this section apply; e.g., we get the familiar formulas $E[I_E] = P(E)$, for an event E, and

$$E[aX + bY] \;=\; aE[X] + b[Y]\,, \tag{A.62}$$

for jointly-distributed integrable random variables X and Y and constants a and b, as in (A.59). This extends to any finite number of random variables, by induction. One of the most important results of probability theory is stated next, without proof.

Theorem A.5. *(Change of Variable Theorem.) If $g : R \to R$ is a measurable function, then*

$$E[g(X)] \;=\; \int_S g(X(\omega))\, P(d\omega) \;=\; \int_{-\infty}^{\infty} g(x)\, P_X(dx)\,, \tag{A.63}$$

where P_X is the distribution of X, defined in (A.30).

Hence, expectations can be computed by integration over the real line. The previous theory is entirely general, and applies equally well to continuous, discrete, and more general random variables.

If X is continuous, then it satisfies (A.41), where the integral should be interpreted as Lebesgue integral over the interval $[a, b]$. It can be shown then that p_X is a density for the distribution P_X with respect to Lebesgue measure. Combining Theorems A.4 and A.5 produces the familiar formula:

$$E[g(X)] \;=\; \int_{-\infty}^{\infty} g(x)\, p_X(x)\, dx\,, \tag{A.64}$$

where the integral is the Lebesgue integral, which reduces to the ordinary integral if the integrand is Riemann-integrable. If $g(x) = x$, one gets the usual definition $E[X] = \int x\, p_X(x)\, dx$.

On the other hand, if X is discrete, then P_X is concentrated on a countable number of points $x_1, x_2, \ldots$, and Thm A.5 produces

$$E[g(X)] \;=\; \sum_{k=1}^{\infty} g(x_k)\, p_X(x_k)\,, \tag{A.65}$$

if the sum is well-defined. If $g(x) = x$, we get the familiar formula $E[X] \;=\; \sum_{k=1}^{\infty} x_k\, p_X(x_k)$.

From now on we assume that random variables are integrable. If $f : R \to R$ is Borel-measurable and *concave* (i.e., f lies at or above a line joining any of its points) then *Jensen's Inequality* is:

$$E[f(X)] \;\leq\; f(E[X])\,. \tag{A.66}$$

It can be shown that X and Y are independent if and only if $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all Borel-measurable functions $f, g : R \to R$. If this condition is satisfied for at least $f(X) = X$ and

$g(Y) = Y$, that is, if $E[XY] = E[X]E[Y]$, then X and Y are said to be *uncorrelated*. Of course, independence implies uncorrelatedness. The converse is only true in special cases; e.g. jointly Gaussian random variables.

Holder's Inequality states that, for $1 < r < \infty$ and $1/r + 1/s = 1$,

$$E[|XY|] \;\leq\; E[|X|^r]^{1/r} E[|Y|^s]^{1/s}. \tag{A.67}$$

The special case $r = s = 2$ results in the *Cauchy-Schwarz Inequality*:

$$E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}\,. \tag{A.68}$$

The expectation of a random variable X is affected by its *probability tails*, given by $F_X(a) = P(X \leq a)$ and $1 - F_X(a) = P(X \geq a)$. If the probability tails on both sides fail to vanish sufficiently fast (X has "fat tails"), then X will not be integrable and $E[X]$ is undefined. The standard example is the Cauchy random variable, with density $p_X(x) = [\pi(1 + x^2)]^{-1}$. For a nonnegative random variable X, there is only one probability tail, the upper tail $P(X > a)$, and there is a simple formula relating $E[X]$ to it:

$$E[X] = \int_0^\infty P(X > x)\, dx\,. \tag{A.69}$$

A small $E[X]$ constrain the upper tail to be thin. This is guaranteed by *Markov's inequality:* if X is a nonnegative random variable,

$$P(X \geq a) \;\leq\; \frac{E[X]}{a}\,, \quad \text{for all } a > 0\,. \tag{A.70}$$

Finally, a particular result that if of interest to our purposes relates an exponentially-vanishing upper tail of a nonnegative random variable to a bound on its expectation.

Lemma A.1. *If X is a non-negative random variable such that $P(X > t) \leq ce^{-at^2}$, for all $t > 0$ and given $a, c > 0$, we have*

$$E[X] \leq \sqrt{\frac{1 + \ln c}{a}}\,. \tag{A.71}$$

Proof. Note that $P(X^2 > t) = P(X > \sqrt{t}) \leq ce^{-at}$. From (A.69) we get:

$$\begin{aligned} E[X^2] \;=\; \int_0^\infty P(X^2 > t)\, dt \;&=\; \int_0^u P(X^2 > t)\, dt + \int_u^\infty P(X^2 > t)\, dt \\ &\leq\; u + \int_u^\infty ce^{-at}\, dt \;=\; u + \frac{c}{a} e^{-au}\,. \end{aligned} \tag{A.72}$$

By direct differentiation, it is easy to verify that the upper bound on the right hand side is minimized at $u = (\ln c)/a$. Substituting this value back into the bound leads to $E[X^2] \leq (1 + \ln c)/a$. The result then follows from the fact that $E[X] \leq \sqrt{E[X^2]}$. $\diamond$

If the second moment exists, the variance $\mathrm{Var}(X)$ of a random variable X is a nonnegative quantity defined by:

$$\mathrm{Var}(X) \;=\; E[(X - E[X])^2] = E[X^2] - (E[X])^2. \tag{A.73}$$

The variance of a random variable can be interpreted as: 1) its "spread" around the mean; 2) a second summary of its distribution (the "scale parameter"); 3) the uncertainty in the prediction of its future value by its expectation.

The following property follows directly from the definition:

$$\mathrm{Var}(aX + c) \;=\; a^2\mathrm{Var}(X)\,. \tag{A.74}$$

A small variance constrains the random variable to be close to its mean with high probability. This follows from *Chebyshev's Inequality*:

$$P(|X - E[X]| \geq \tau) \leq \frac{\mathrm{Var}(X)}{\tau^2}\,, \quad \text{for all } \tau > 0\,. \tag{A.75}$$

Chebyshev's inequality follows directly from the application of Markov's Inequality (A.70) to the random variable $|X - E[X]|^2$ with $a = \tau^2$.

Expectation has the linearity property, so that, given any pair of jointly distributed random variables X and Y, it is always true that $E[X + Y] = E[X] + E[Y]$ (provided that all expectations exist). However, it is not always true that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. In order to investigate this issue, it is necessary to introduce the *covariance* between X and Y:

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]\,. \tag{A.76}$$

If $\mathrm{Cov}(X, Y) > 0$ then X and Y are *positively correlated*; otherwise, they are *negatively correlated*. Clearly, X and Y are uncorrelated if and only if $\mathrm{Cov}(X, Y) = 0$. Clearly, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. In addition, $\mathrm{Cov}(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, Y_j)$.

Now, it follows directly from the definition of variance that

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)\,. \tag{A.77}$$

This can be extended to any number of random variables by induction:

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \;=\; \sum_{i=1}^n \mathrm{Var}(X_i) \;+\; 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)\,. \tag{A.78}$$

Hence, the variance is distributive over sums if all variables are *pairwise uncorrelated*. o It follows directly from the Cauchy-Schwarz Inequality (A.68) that $|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$. Therefore, the covariance can be normalized to be in the interval $[-1, 1]$ thus:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}, \tag{A.79}$$

with $-1 \leq \rho(X,Y) \leq 1$. This is called the correlation coefficient between X and Y. The closer $|\rho|$ is to 1, the tighter is the relationship between X and Y. The limiting case where $\rho(X,Y) = \pm 1$ occurs if and only if $Y = a \pm bX$, i.e., X and Y are perfectly related to each other through a linear (affine) relationship. For this reason, $\rho(X,Y)$ is sometimes called the linear correlation coefficient between X and Y; it does not respond to nonlinear relationships.

Conditional expectation allows the prediction of the value of a random variable given the *observed* value of the other, i.e., prediction given data, while conditional variance yields the uncertainty of that prediction.

If X and Y are jointly continuous random variables and the conditional density $p_{X|Y}(x \mid y)$ is well defined for $Y = y$, then the conditional expectation of X *given* $Y = y$ is:

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x\, p_{X|Y}(x \mid y)\, dx \tag{A.80}$$

with a similar definition for discrete random variables using conditional PMFs.

The conditional variance of X given $Y = y$ is defined using conditional expectation as:

$$\mathrm{Var}(X \mid Y = y) \;=\; E[(X - E[X \mid Y = y])^2 \mid Y = y] \;=\; E[X^2 \mid Y = y] - (E[X \mid Y = y])^2\,. \tag{A.81}$$

Most of the properties of expectation and variance apply without modification to conditional expectations and conditional variances, respectively. For example, $E[\sum_{i=1}^{n} X_i \mid Y = y] = \sum_{i=1}^{n} E[X_i \mid Y = y]$ and $\mathrm{Var}(aX + c \mid Y = y) = a^2 \mathrm{Var}(X \mid Y = y)$.

Now, both $E[X \mid Y = y]$ and $\mathrm{Var}(X \mid Y = y)$ are deterministic quantities for each value of $Y = y$ (just as the ordinary expectation and variance are). But if the specific value $Y = y$ is not specified and allowed to vary, then we can look at $E[X \mid Y]$ and $\mathrm{Var}(X \mid Y)$ as functions of the random variable Y, and therefore, random variables themselves. The reasons why these are valid random variables are nontrivial and beyond the scope of this review.

One can show that the expectation of the random variable $E[X \mid Y]$ is precisely $E[X]$:

$$E[E[X \mid Y]] = E[X]\,. \tag{A.82}$$

An equivalent statement is:

$$E[X] = \int_{-\infty}^{\infty} E[X \mid Y = y]\, p(y)\, dy\,, \tag{A.83}$$

with a similar expression in the discrete case. Paraphrasing the Law of Total Probability (A.26), the previous equation might be called the *Law of Total Expectation*.

On the other hand, it is not the case that $\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)]$. The answer is slightly more complicated:

$$\mathrm{Var}(X) \;=\; E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y])\,. \tag{A.84}$$

This is known as the *Conditional Variance Formula*. It is an "analysis of variance" formula, as it breaks down the total variance of X into a "within-rows" component and an "across-rows" component. One might call this the *Law of Total Variance*. This formula plays a key role in Chapter 7.

Now, suppose one is interested in predicting the value of a random variable Y using a predictor $\hat{Y}$. One would like $\hat{Y}$ to be optimal according to some criterion. The criterion most widely used is the *mean-square error:*

$$\text{MSE} \;=\; E[(Y-\hat{Y})^2]\,. \tag{A.85}$$

It can be shown easily that the minimum mean-square error (MMSE) estimator is simply the mean: $\hat{Y}^* = E[Y]$. This is a constant estimator, since no data are available. Clearly, the MSE of $\hat{Y}^*$ is simply the variance of Y. Therefore, the best one can do in the absence of any extra information is to predict the mean $E[Y]$, with an uncertainty equal to the variance $\text{Var}(Y)$.

If $\text{Var}(Y)$ is very small, i.e., if there were very small uncertainty in Y to begin with, then $E[Y]$ could actually be an acceptable estimator. In practice, this is rarely the case. Therefore, observations on an auxiliary random variable X (i.e., *data*) are sought to improve prediction. Naturally, it is known (or hoped) that X and Y are not independent, otherwise no improvement over the constant estimator is possible. One defines the *conditional MSE* of a data-dependent estimator $\hat{Y} = h(X)$ as

$$\text{MSE}\,(X) \;=\; E[(Y-h(X))^2 \mid X]\,. \tag{A.86}$$

By taking expectation over X, one obtains the unconditional MSE: $E[(Y-h(X))^2]$. The conditional MSE is often the most important one in practice, since it concerns the particular data at hand, while the unconditional MSE is data-independent and used to compare the performance of different predictors. Regardless, the MMSE estimator in *both cases* is the conditional mean $h^*(X) = E[Y \mid X]$, as shown in Chapter 11. This is one of the most important results in supervised learning. The *posterior-probability function* $\eta(x) = E[Y \mid X = x]$ is the optimal regression of Y on X. This is not in general the optimal estimator if Y is discrete; e.g., in the case of classification. This is because $\eta(X)$ may not be in the range of values taken by Y, so it does not define a valid estimator. It is shown in Chapter 2 that one needs to threshold $\eta(x)$ at $1/2$ to obtain the optimal estimator (optimal classifier) in the case $Y \in \{0, 1\}$.

A1.7 Vector Random Variables

The previous theory can be extended to vector random variables, or *random vectors*, defined on a probability space $(S, \mathcal{F}, P)$. A random vector is a Borel-measurable function $\mathbf{X} : S \to R^d$, with a probability distribution $P_{\mathbf{X}}$ defined on $(R^d, \mathcal{B}^d)$. The components of the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ are jointly-distributed random variables X_i on $(S, \mathcal{F}, P)$, for $i = 1, \ldots, d$.

The expected value of $\mathbf{X}$ is the vector of expected values of the components, if they exist:

$$E[\mathbf{X}] = \begin{bmatrix} E[X_1] \\ \cdots \\ E[X_d] \end{bmatrix}. \tag{A.87}$$

The second moments of a random vector are contained in the $d \times d$ *covariance matrix*:

$$\Sigma = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T], \tag{A.88}$$

where$\Sigma_{ii} = \mathrm{Var}(X_i)$ and$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, for $i, j = 1, \ldots, d$, and the expectation of the matrix is defined as the matrix of the expected values of its components, assuming they exist. The covariance matrix is real symmetric and thus diagonalizable:

$$\Sigma = UDU^T, \tag{A.89}$$

where U is the orthogonal matrix of eigenvectors and D is the diagonal matrix of eigenvalues (a review of basic matrix theory facts is given in Section A2). All eigenvalues are nonnegative ($\Sigma$is *positive semi-definite*). In fact, except for "degenerate" cases, all eigenvalues are positive, and soΣ is invertible ($\Sigma$is said to be *positive definite* in this case).

It is easy to check that the random vector

$$\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}) = D^{-\frac{1}{2}}U^T(\mathbf{X} - \boldsymbol{\mu}) \tag{A.90}$$

has zero mean and covariance matrix $\mathbf{I}_d$ (so that all components of $\mathbf{Y}$ are zero-mean, unit-variance, and uncorrelated). This is called *whitening or the Mahalanobis transformation*.

Given n *independent and identically-distributed* (i.i.d.) sample observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the random vector $\mathbf{X}$, then the maximum-likelihood estimator of $\boldsymbol{\mu} = E[\mathbf{X}]$, known as the *sample mean*, is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \,. \tag{A.91}$$

It can be shown that this estimator is *unbiased* (that is, $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$) and *consistent* (that is, $\hat{\boldsymbol{\mu}}$ converges in probability to $\boldsymbol{\mu}$ as $n \to \infty$; see Section A1.8 and Theorem A.12). On the other hand, the *sample covariance* estimator is given by:

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^T \,. \tag{A.92}$$

This is an unbiased and consistent estimator ofΣ.

The multivariate Gaussian distribution is probably the most important probability distribution in Engineering and Science. The random vector $\mathbf{X}$ has a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrixΣ(assumingΣinvertible) if its density is given by

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{A.93}$$

We write $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$.

The multivariate Gaussian has ellipsoidal contours of constant density of the form

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2, \quad c > 0\,. \tag{A.94}$$

The axes of the ellipsoids are given by the eigenvectors of$\Sigma$and the length of the axes are proportional to its eigenvalues. In the case$\Sigma = \sigma^2 I_d$, where I_d denotes the $d \times d$ identity matrix, the contours are spherical with center at $\boldsymbol{\mu}$. This can be seen by substituting$\Sigma = \sigma^2 I_d$ in (A.94), which leads to the following equation for the contours:

$$||\mathbf{x}-\boldsymbol{\mu}||^2 = r^2, \quad r > 0\,, \tag{A.95}$$

If $d = 1$, one gets the univariate Gaussian distribution $X \sim \mathcal{N}(\mu,\sigma^2)$. With $\mu = 0$ and $\sigma = 1$, the CDF of X is given by

$$P(X \leq x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-\frac{u^2}{2}} du\,. \tag{A.96}$$

It is clear that the function$\Phi(\cdot)$ satisfies the property$\Phi(-x) = 1 - \Phi(x)$.

The following are useful properties of a multivariate Gaussian random vector $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$:

G1. The density of each component X_i is univariate gaussian $\mathcal{N}(\mu_i, \Sigma_{ii})$.

G2. The components of $\mathbf{X}$ are independent *if and only if* they are uncorrelated, i.e.,$\Sigma$is a diagonal matrix.

G3. The whitening transformation $\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X}-\boldsymbol{\mu})$ produces a multivariate gaussian $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}_p)$ (so that all components of $\mathbf{Y}$ are zero-mean, unit-variance, and uncorrelated Gaussian random variables).

G4. In general, if $\mathbf{A}$ is a nonsingular $p \times p$ matrix and $\mathbf{c}$ is a p-vector, then $\mathbf{Y} = \mathbf{AX} + \mathbf{c} \sim N_p(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\Sigma\mathbf{A}^T)$.

G5. The random vectors $\mathbf{AX}$ and $\mathbf{BX}$ are independent iff $\mathbf{A}\Sigma\mathbf{B}^T = 0$.

G6. If $\mathbf{Y}$ and $\mathbf{X}$ are jointly multivariate Gaussian, then the distribution of $\mathbf{Y}$ given $\mathbf{X}$ is again multivariate Gaussian.

G7. The best MMSE predictor $E[\mathbf{Y} \mid \mathbf{X}]$ is a linear function of $\mathbf{X}$.

A1.8 Convergence of Random Sequences

It is often necessary in pattern recognition and machine learning to investigate the long-term behavior of random sequences, such as the sequence of true or estimated classification error rates indexed by sample size. In this section and the next, we review basic results about convergence of random sequences. We consider only the case of real-valued random variables, but nearly all the definitions and results can be directly extended to random vectors, with the appropriate modifications.

A *random sequence* $\{X_n;\, n = 1, 2, \ldots\}$ is a sequence of random variables. The standard modes of convergence for random sequences are:

1. *"Sure"* convergence: $X_n \to X$ surely if for all outcomes $\omega \in S$ in the sample space one has $\lim_{n\to\infty} X_n(\omega) = X(\omega)$.

2. *Almost-sure (a.s.) convergence* or *convergence with probability 1*: $X_n \xrightarrow{a.s.} X$ if pointwise converge fails only for an event of probability zero, i.e.:

$$P\left(\left\{\omega \in S \;\middle|\; \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) \;=\; 1\,. \tag{A.97}$$

3. *L^p-convergence*: $X_n \to X$ in L^p, for $p > 0$, also denoted by $X_n \xrightarrow{L^p} X$, if $E[|X_n|^p] < \infty$ for $n = 1, 2, \ldots$, $E[|X|^p] < \infty$, and:

$$\lim_{n\to\infty} E[|X_n - X|^p] \;=\; 0\,. \tag{A.98}$$

   The special case of L^2 convergence is also called *mean-square* (m.s.) convergence.

4. *Convergence in probability*: $X_n \to X$ in probability, also denoted by $X_n \xrightarrow{P} X$, if the "probability of error" converges to zero:

$$\lim_{n\to\infty} P(|X_n - X| > \tau) = 0\,, \quad \text{for all } \tau > 0\,. \tag{A.99}$$

5. *Convergence in Distribution* : $X_n \to X$ in distribution, also denoted by $X_n \xrightarrow{D} X$, if the corresponding CDFs converge:

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a)\,, \tag{A.100}$$

   at all points $a \in R$ where F_X is continuous.

We state, without proof, the relationships among the various modes of convergence:

$$\left.\begin{array}{c}\text{sure} \Rightarrow \text{almost-sure} \\ L^p\end{array}\right\} \Rightarrow \text{probability} \Rightarrow \text{distribution}\,. \tag{A.101}$$

Hence, sure convergence is the strongest mode of convergence and convergence in distribution is the weakest. However, sure convergence is unnecessarily demanding, and almost-sure convergence is the strongest mode of convergence employed. On the other hand, convergence is distribution is really convergence of CDFs, and does not have all the properties one expects from convergence. For example, it can be shown that convergence X_n to X and Y_n to Y in distribution does not imply in general that $X_n + Y_n$ converges to $X + Y$ in distribution, whereas this would be true for convergence almost surely, in L^p, and in probability [Chung, 1974].

To show consistency of parametric classification rules (see Chapters 3 and 4), an essential fact about convergence with probability 1 and in probability is that, similarly to ordinary convergence, they are preserved by application of continuous functions. The following result is stated without proof.

Theorem A.6. *(Continuous Mapping Theorem.) If $f : R \to R$ is continuous a.e. with respect to X, i.e. $P(X \in C) = 1$, where C is the set of points of continuity of f, then*

(i) $X_n \xrightarrow{a.s.} X$ *implies that* $f(X_n) \xrightarrow{a.s.} f(X)$.

(ii) $X_n \xrightarrow{P} X$ *implies that* $f(X_n) \xrightarrow{P} f(X)$.

(iii) $X_n \xrightarrow{D} X$ *implies that* $f(X_n) \xrightarrow{D} f(X)$.

A special case of interest is $X = c$, i.e., the distribution of X is a point mass at c. In this case, the continuous mapping theorem requires f to be merely continuous at c.

The following classical result is stated here without proof.

Theorem A.7. *(Dominated Convergence Theorem.) If there is an integrable random variable Y, i.e., $E[|Y|] < \infty$, with $P(|X_n| \leq Y) = 1$, for $n = 1, 2, \ldots$, then $X_n \xrightarrow{P} X$ implies that $E[X_n] \to E[X]$.*

The next result provides a common way to show strong consistency (e.g., see Chapter 7). It is a consequence of the First Borel-Cantelli Lemma, and it indicates that converge with probability 1 is in a sense a "fast" form of convergence in probability.

Theorem A.8. *If, for all $\tau > 0$, $P(|X_n - X| > \tau) \to 0$ fast enough to obtain*

$$\sum_{n=1}^{\infty} P(|X_n - X| > \tau) < \infty \,, \tag{A.102}$$

then $X_n \xrightarrow{a.s.} X$.

Proof. First notice that a sample sequence $X_n(\omega)$ fails to converge to $X(\omega)$ if and only if there is a $\tau > 0$ such that $|X_n(\omega)) - X(\omega)| > \tau$ infinitely often as $n \to \infty$. Hence, $X_n \to X$ a.s. if and only

if $P(|X_n - X| > \tau)\ i.o.) = 0$, for all $\tau > 0$. The result then follows from the First Borel-Cantelli Lemma (see Thm. A.1). $\diamond$

The previous result implies that convergence in probability can produce convergence with probability 1 along a subsequence, obtained by "downsampling" the original sequence, as shown next.

Theorem A.9. *If $X_n \xrightarrow{P} X$, then there is an increasing sequence of indices n_k such that $X_{n_k} \xrightarrow{a.s.} X$.*

Proof. Since $P(|X_n - X| > \tau) \to 0$, for all $\tau > 0$, we can pick an increasing sequence of indices n_k such that $P(|X_{n_k} - X| > 1/k) \leq 2^{-k}$. Given any $\tau > 0$, pick k_τ such that $1/k_\tau < \tau$. We have

$$\sum_{k=k_\tau}^{\infty} P(|X_{n_k} - X| > \tau) \leq \sum_{k=k_\tau}^{\infty} P(|X_{n_k} - X| > 1/k) \leq \sum_{k=k_\tau}^{\infty} 2^{-k} < \infty\,, \tag{A.103}$$

so that $X_{n_k} \xrightarrow{a.s.} X$ by Theorem A.8. $\diamond$

The previous theorem provides a criterion to *disprove* convergence $X_n \to X$ in probability: it is enough to show that there is no subsequence that converges to X with probability 1. This criterion is used in Chapter 4 (see Example 4.4).

Notice also that if X_n is monotone and $P(|X_n - X| > \tau) \to 0$, then $P(|X_n - X| > \tau)\ i.o.) = 0$. Hence, if X_n is monotone, $X_n \to X$ in probability if and only if $X_n \to X$ with probability 1 (see the proof of Thm. A.8).

Stronger relations among the modes of convergence hold in special cases. In particular, we prove below that L^p convergence and convergence in probability are equivalent if the random sequence $\{X_n;\ n = 1, 2, \ldots\}$ is *uniformly bounded*, i.e., if there exists a finite $K > 0$, *which does not depend on n*, such that

$$|X_n| \leq K\,, \text{with probability 1, for all } n = 1, 2, \ldots \tag{A.104}$$

meaning that $P(|X_n| < K) = 1$, for all $n = 1, 2, \ldots$ The classification error rate sequence $\{\varepsilon_n;\ n = 1, 2, \ldots\}$ is an example of uniformly bounded random sequence, with $K = 1$, therefore this is an important topic for our purposes. We have the following theorem.

Theorem A.10. *Let $\{X_n;\ n = 1, 2, \ldots\}$ be a uniformly bounded random sequence. The following statements are equivalent.*

(1) $X_n \xrightarrow{L^p} X$, for some $p > 0$.

(2) $X_n \xrightarrow{L^q} X$, for all $q > 0$.

(3) $X_n \xrightarrow{P} X$.

Proof. First note that we can assume without loss of generality that $X = 0$, since $X_n \to X$ if and only if $X_n - X \to 0$, and $X_n - X$ is also uniformly bounded, with $E[|X_n - X|^p] < \infty$. Showing that (1) $\Leftrightarrow$ (2) requires showing that $X_n \to 0$ in L^p, for some $p > 0$ implies that $X_n \to 0$ in L^q, for all $q > 0$. First observe that $E[|X_n|^q] \le E[K^q] = K^q < \infty$, for all $q > 0$. If $q > p$, the result is immediate. Let $0 < q < p$. With $X = X_n^q$, $Y = 1$ and $r = p/q$, Holder's Inequality (A.67) yields

$$E[|X_n|^q] \;\le\; E[|X_n|^p]^{q/p}\,. \tag{A.105}$$

Hence, if $E[|X_n|^p] \to 0$, then $E[|X_n|^q] \to 0$, proving the assertion. To show that (2) $\Leftrightarrow$ (3), first we show the direct implication by writing Markov's Inequality (A.70) with $X = |X_n|^p$ and $a = \tau^p$:

$$P(|X_n| \ge \tau) \le \frac{E[|X_n|^p]}{\tau^p}\,, \quad \text{for all } \tau > 0\,. \tag{A.106}$$

The right-hand side goes to 0 by hypothesis, and thus so does the left-hand side, which is equivalent to (A.99) with $X = 0$. To show the reverse implication, write

$$E[|X_n|^p] \;=\; E[|X_n|^p I_{|X_n|<\tau}] + E[|X_n|^p I_{|X_n|\ge\tau}] \;\le\; \tau^p + K^p P(|X_n| \ge \tau)\,. \tag{A.107}$$

By assumption, $P(|X_n| \ge \tau) \to 0$, for all $\tau > 0$, so that $\lim E[|X_n|^p] \le \tau^p$. Since $\tau > 0$ is arbitrary, this establishes the desired result. $\diamond$

The previous theorem implies that, for uniformly bounded random sequences, the relationships among the modes of convergence become:

$$\text{sure} \Rightarrow \text{almost-sure} \Rightarrow \left\{ \begin{array}{c} L^p \\ \text{probability} \end{array} \right\} \Rightarrow \text{distribution} \tag{A.108}$$

As a simple corollary of Theorem A.10, we have the following useful result, which is also a corollary of Theorem A.7.

Theorem A.11. *(Bounded Convergence Theorem.) If $\{X_n;\, n = 1, 2, \ldots\}$ is a uniformly bounded random sequence and $X_n \xrightarrow{P} X$, then $E[X_n] \to E[X]$.*

Proof. From the previous theorem, $X_n \xrightarrow{L^1} X$, i.e., $E[|X_n - X|] \to 0$. But $|E[X_n - X]| \le E[|X_n - X|]$, hence $|E[X_n - X]| \to 0$ and $E[X_n - X] \to 0$. $\diamond$

Example A.1. To illustrate these concepts, consider a sequence of independent binary random variables $X_1, X_2, \ldots$ that take on values in $\{0, 1\}$ such that

$$P(\{X_n = 1\}) = \frac{1}{n}, \quad n = 1, 2, \ldots \tag{A.109}$$

Then $X_n \xrightarrow{P} 0$, since $P(X_n > \tau) \to 0$, for every $\tau > 0$. By Theorem A.10, $X_n \xrightarrow{L^p} 0$ as well. However, X_n does not converge to 0 with probability 1. Indeed,

$$\sum_{n=1}^{\infty} P(\{X_n = 1\}) \;=\; \sum_{n=1}^{\infty} P(\{X_n = 0\}) \;=\; \infty\,, \tag{A.110}$$

and it follows from the 2nd Borel-Cantelli lemma that

$$P([\{X_n = 1\}\; i.o.]) \;=\; P([\{X_n = 0\}\; i.o.]) = 1, \tag{A.111}$$

so that X_n does not converge with probability 1. However, if convergence of the probabilities to zero is faster, e.g.

$$P(\{X_n = 1\}) = \frac{1}{n^2}, \quad n = 1, 2, \ldots \tag{A.112}$$

then $\sum_{n=1}^{\infty} P(\{X_n = 1\}) < \infty$ and Theorem A.8 ensures that X_n converges to 0 with probability 1. $\diamond$

In the previous example, note that, with $P(X_n = 1) = 1/n$, the probability of observing a 1 becomes infinitesimally small as $n \to \infty$, so the sequence consists, for all practice purposes, of all zeros for large enough n. Convergence in probability and in L^p of X_n to 0 agrees with this fact, but the lack of convergence with probability 1 does not. This is an indication that almost-sure convergence may be too stringent a criterion to be useful in practice, and convergence in probability and in L^p (assuming boundedness) may be enough. For example, this is the case in most signal processing applications, where L^2 is the criterion of choice. More generally, Engineering applications usually concern average performance and rates of failure.

A1.9 Asymptotic Theorems

The classical asymptotic theorems in probability theory are the Law of Large Numbers and the Central Limit Theorem, the proofs of which can be found, for example, in Chung [1974].

Theorem A.12. *(Law of Large Numbers.) Given an i.i.d. random sequence* $\{X_n;\, n = 1, 2, \ldots\}$ *with common finite mean* μ*,*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu\,. \tag{A.113}$$

Theorem A.13. *(Central Limit Theorem.) Given an i.i.d. random sequence* $\{X_n;\, n = 1, 2, \ldots\}$ *with common finite mean* μ *and common finite variance* σ^2*,*

$$\frac{1}{\sigma\sqrt{n}} \left(\sum_{i=1}^{n} X_i - n\mu \right) \xrightarrow{D} \mathcal{N}(0, 1)\,. \tag{A.114}$$

The previous asymptotic theorems concern the behavior of a sum of n random variables as n approach infinity. It is also useful to have an idea of how partial sums differ from expected values for finite n. This issue is addressed by the so-called *concentration inequalities*, the most famous of which is Hoeffding's inequality, derived in Hoeffding [1963].

Theorem A.14. *(Hoeffding's Inequality.) Given independent (not necessarily identically-distributed) random variables $W_1, \ldots, W_n$ such that $P(a \leq W_i \leq b) = 1$, for $i = 1, \ldots, n$, the sum $Z_n = \sum_{i=1}^n W_i$ satisfies*

$$P(|Z_n - E[Z_n]| \geq \tau) \leq 2e^{-\frac{2\tau^2}{n(a-b)^2}}, \quad \text{for all } \tau > 0\,. \tag{A.115}$$

A2 Basic Matrix Theory

The material in this section is a summary of concepts and results from main matrix theory that are useful in the text. For an in-depth treatment, see Horn and Johnson [1990].

We assume that the reader is familiar with the concepts of vector, matrix, matrix product, transpose, determinant, and matrix inverse. We say that a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is *linearly dependent* if the equation

$$a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n = 0 \tag{A.116}$$

is satisfied for coefficients $a_1, \ldots, a_n$ that are not all zero. In other words, some of the vectors can be written as a linear combination of other vectors. If a set of vectors is not linearly dependent, then it is said to be *linearly independent*.

The *rank* of a matrix $A_{m\times n}$ is the largest number of columns of A that form a linearly independent set. This must be equal to the maximum number of rows that form a linearly independent set (row rank = column rank). A square matrix $A_{n\times n}$ is *nonsingular* if the inverse A^{-1} exists, or equivalently, the determinant $|A|$ is nonzero. The following are useful facts:

- $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(AA^T) = \text{rank}(A^TA)$, where A^T denotes matrix transpose.
- $\text{rank}(A_{m\times n}) \leq \min\{m, n\}$. If equality is achieved, A is said to be *full-rank*.
- $A_{n\times n}$ is nonsingular if and only if $\text{rank}(A) = n$, i.e., A is full-rank. By the definition of rank, this means that the system of equations $A\mathbf{x} = 0$ has a unique solution $\mathbf{x} = 0$..
- If $B_{m\times m}$ is nonsingular then $\text{rank}(BA_{m\times n}) = \text{rank}(A)$ (multiplication by a nonsingular matrix preserves rank).
- $\text{rank}(A_{m\times n}) = \text{rank}(B_{m\times n})$ if and only if there are nonsingular matrices $X_{m\times m}$ and $Y_{n\times n}$ such that $B = XAY$.

- If $\text{rank}(A_{m\times n}) = k$, then there is a nonsingular matrix $B_{k\times k}$ and matrices $X_{m\times k}$ and $Y_{k\times n}$ such that $A = XBY$.
- As a corollary from the previous fact, $A_{m\times n}$ is a *rank-1 matrix* if A is a product of two vectors, $A = \mathbf{x}\mathbf{y}^T$, where the lengths of $\mathbf{x}$ and $\mathbf{y}$ are m and n, respectively.

An *eigenvalue* λ of a square matrix $A_{n\times n}$ is a solution of the equation

$$A\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{x} \neq 0, \tag{A.117}$$

in which case $\mathbf{x}$ is an *eigenvector* of A associated with λ. Complex λ and $\mathbf{x}$ are allowed. The following are useful facts:

- The eigenvalues of A and A^T are the same.
- If A is real symmetric, then all its eigenvalues are real.
- Since A is singular if and only if $A\mathbf{x} = 0$ with nonzero $\mathbf{x}$, we conclude that A is singular if and only if it has a zero eigenvalue.

From (A.117), λ is an eigenvalue if and only if $(A - \lambda I_n)\mathbf{x} = 0$ with nonzero $\mathbf{x}$. From previous facts, we conclude that $A - \lambda I_n$ is singular, that is, $|A - \lambda I_n| = 0$. But $p(\lambda) = |A - \lambda I_n|$ is a polynomial of degree n, which thus has exactly n roots (allowing for multiplicity), so we have proved the following useful fact.

Theorem A.15. *Any square matrix $A_{n\times n}$ has exactly n (possibly complex) eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, which are the roots of the* characteristic polynomial $p(\lambda) = |A - \lambda I_n|$.

If A is a diagonal matrix, then the eigenvalues are clearly the elements in its diagonal, so that $\text{Trace}(A) = \sum_{i=1}^n \lambda_i$ and $|A| = \prod_{i=1}^n \lambda_i$. It is a remarkable fact that it is still true that $\text{Trace}(A) = \sum_{i=1}^n \lambda_i$ and $|A| = \prod_{i=1}^n \lambda_i$ for any, not necessarily diagonal, square matrix A.

Matrix $B_{n\times n}$ is *similar* to matrix $A_{n\times n}$ if there is a nonsingular matrix $S_{n\times n}$ such that

$$B = S^{-1}AS. \tag{A.118}$$

It is easy to show that if A and B are similar, they have the same characteristic polynomial, and therefore the same set of eigenvalues (however, having the same set of eigenvalues is not sufficient for similarity).

Matrix A is said to be *diagonalizable* if it is similar to a diagonal matrix D. Since similarity preserves the characteristic polynomial, the eigenvalues of A are equal to the elements in the diagonal of D. The following theorem is not difficult to prove.

Theorem A.16. *A matrix $A_{n\times n}$ is diagonalizable if and only if it has a set of n linearly independent eigenvectors.*

A real-valued matrix $U_{n\times n}$ is said to be *orthogonal* if $U^TU = UU^T = I_n$, i.e., $U^{-1} = U^T$. Clearly, this happens if and only if the columns (and rows) of U are a set of unit-norm orthogonal vectors in R^n. Matrix $A_{n\times n}$ is said to be *orthogonally diagonalizable* if it is diagonalizable by an orthogonal matrix $U_{n\times n}$, i.e., $A = U^TDU$, where D is diagonal. Since

The following theorem, stated without proof, is one of the most important results in matrix theory.

Theorem A.17. *(Spectral Theorem.) If A is real symmetric, then it is orthogonally diagonalizable.*

Therefore, of A is real symmetric, we can write $A = U^T\Lambda U$ and$\Lambda = UAU^T$, where λ is a diagonal matrix containing the n eigenvalues of A on its diagonal. Furthermore, $UA = \Lambda U$, and thus the $i - the$ column of U is the eigenvector of A associated with the eigenvalue in the i-the position of the diagonal ofΛ, for $i = 1, \ldots, n$.

A real symmetric matrix $A_{n\times n}$ is said to be *positive definite* if

$$\mathbf{x}^TA\mathbf{x} > 0\,, \quad \text{for all } \mathbf{x} \neq 0\,. \tag{A.119}$$

If the condition is relaxed to $\mathbf{x}^TA\mathbf{x} \geq 0$, then A is said to be *positive semi definite.* As we mentioned in the text, a covariance matrix is always at least positive semi-definite.

The following theorem is not difficult to prove.

Theorem A.18. *A real symmetric matrix A is positive definite if and only if all its eigenvalues are positive. It is positive semidefinite if and only if all eigenvalues are nonnegative.*

In particular, a positive definite matrix A is nonsingular. Another useful fact is that A is positive definite if and only if there is a nonsingular matrix C such that $A = CC^T$.

A3 Basic Lagrange-Multiplier Optimization

In this section we review results from Lagrange Multiplier theory that are needed in Section 6.1.1. For simplicity, we consider only minimization with inequality constraints, which is the case of the linear SVM optimization problems (6.6) and (6.20). Our presentation follows largely Chapter 5 of Boyd and Vandenberghe [2004], with some elements from Chapters 5 and 6 of Bertsekas [1995].

Consider the general (not necessarily convex) optimization problem:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0\,, \quad i = 1, \ldots, n\,. \end{aligned} \tag{A.120}$$

where all functions are defined on R^d.

The *primal Lagrangian functional* is defined as

$$L_P(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x})\,, \tag{A.121}$$

where λ_i is the *Lagrange multiplier* associated with constraint $g_i(\mathbf{x}) \leq 0$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$.

The *dual Lagrangian functional* is defined as:

$$L_D(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x} \in R^d} \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) \right). \tag{A.122}$$

Using the properties of infimum, we have

$$\begin{aligned} L_D(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2) &= \inf_{\mathbf{x} \in R^d} \left(f(\mathbf{x}) + \sum_{i=1}^{n} (\alpha\lambda_{1,i} + (1-\alpha)\lambda_{2,i}) g_i(\mathbf{x}) \right) \\ &= \inf_{\mathbf{x} \in R^d} \left(\alpha \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{1,i} g_i(\mathbf{x}) \right) + (1-\alpha) \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{2,i} g_i(\mathbf{x}) \right) \right) \\ &\geq \alpha \inf_{\mathbf{x} \in R^d} \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{1,i} g_i(\mathbf{x}) \right) + (1-\alpha) \inf_{\mathbf{x} \in R^d} \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{2,i} g_i(\mathbf{x}) \right) \\ &= \alpha L_D(\boldsymbol{\lambda}_1) + (1-\alpha) L_D(\boldsymbol{\lambda}_2)\,, \end{aligned} \tag{A.123}$$

for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in R^n$ and $0 \leq \alpha \leq 1$. The dual Lagrangian functional $L_D(\boldsymbol{\lambda})$ is therefore a *concave* function. Furthermore, for all $\mathbf{x} \in F$, where F is the feasible region of (A.120), and $\boldsymbol{\lambda} \geq 0$,

$$L_P(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) \leq f(\mathbf{x})\,, \tag{A.124}$$

since $g_i(\mathbf{x}) \leq 0$, for $i = 1, \ldots, n$. It follows that

$$L_D(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}) \leq \inf_{\mathbf{x} \in F} f(\mathbf{x}) = f(\mathbf{x}^*)\,, \quad \text{for all } \boldsymbol{\lambda} \geq 0\,, \tag{A.125}$$

showing that $L_D(\boldsymbol{\lambda})$ is a lower bound on $f(\mathbf{x}^*)$, whenever $\boldsymbol{\lambda} \geq 0$.

The natural next step is to maximize this lower bound. This leads to the *dual* optimization problem:

$$\begin{aligned} \max \quad & L_D(\boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \geq 0\,. \end{aligned} \tag{A.126}$$

Since the cost $L_D(\boldsymbol{\lambda})$ is concave (as shown previously) and the feasible region is a convex set, this is a convex optimization problem, for which there are efficient solution methods. This is true whether or not the original problem (A.120) is convex.

If $\boldsymbol{\lambda}^*$ is a solution of (A.126), then it follows from (A.125) that $L_D(\boldsymbol{\lambda}^*) \leq f(\mathbf{x}^*)$, which is known as the *weak duality* property. If equality is achieved,

$$L_D(\boldsymbol{\lambda}^*) \;=\; f(\mathbf{x}^*)\,, \tag{A.127}$$

then the problem is said to satisfy the *strong duality* property. This property is not always satisfied, but there are several sets of conditions, called *constraint qualifications*, that ensure strong duality. For convex optimization problems with affine constraints, such as the linear SVM optimization problems (6.6) and (6.20), a simple constraint qualification condition, known as *Slater's condition*, guarantees strong duality as long as the feasible region is nonempty.

The point $(\bar{\mathbf{w}}, \bar{\mathbf{z}})$, where $\bar{\mathbf{w}} \in W$ and $\bar{\mathbf{z}} \in Z$, is a *saddle point* of a function h defined on $W \times Z$ if

$$h(\bar{\mathbf{y}}, \bar{\mathbf{z}}) \;=\; \inf_{\mathbf{w} \in W} h(\mathbf{w}, \bar{\mathbf{z}}) \quad \text{and} \quad h(\bar{\mathbf{y}}, \bar{\mathbf{z}}) \;=\; \sup_{\mathbf{z} \in Z} h(\bar{\mathbf{w}}, \mathbf{z})\,. \tag{A.128}$$

Under strong duality,

$$\begin{aligned} f(\mathbf{x}^*) \;=\; L_D(\boldsymbol{\lambda}^*) \;&=\; \inf_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}^*) \;=\; \inf_{\mathbf{x} \in R^d} \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i^* g_i(\mathbf{x}) \right) \\ &\leq\; L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;=\; f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* g_i(\mathbf{x}^*) \;\leq\; f(\mathbf{x}^*)\,. \end{aligned} \tag{A.129}$$

The first inequality follows from the definition of inf, whereas the second inequality follows from the facts that $\lambda_i^* \geq 0$ and $g_i(\mathbf{x}^*) \leq 0$, for $i = 1, \ldots, n$. It follows from (A.129) that both inequalities hold with equality. In particular,

$$L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;=\; \inf_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}^*)\,. \tag{A.130}$$

On the other hand, it is always true that

$$\sup_{\boldsymbol{\lambda} \geq 0} L_P(\mathbf{x}^*, \boldsymbol{\lambda}) \;=\; \sup_{\boldsymbol{\lambda} \geq 0} \left(f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* g_i(\mathbf{x}^*) \right) \;=\; f(\mathbf{x}^*)\,, \tag{A.131}$$

because $g_i(\mathbf{x}^*) \leq 0$, for $i = 1, \ldots, n$, so that $f(\mathbf{x}^*)$ maximizes $L_P(\mathbf{x}^*, \boldsymbol{\lambda})$ at $\boldsymbol{\lambda} = 0$. With the extra condition of strong duality, we have from (A.129) that $f(\mathbf{x}^*) = L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, so we obtain

$$L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;=\; \sup_{\boldsymbol{\lambda} \geq 0} L_P(\mathbf{x}^*, \boldsymbol{\lambda})\,. \tag{A.132}$$

It follows from (A.130) and (A.132) that strong duality implies that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a saddle point of $L_p(\mathbf{x}, \boldsymbol{\lambda})$. It follows immediately from the general relations

$$f(\mathbf{x}^*) \;=\; \sup_{\boldsymbol{\lambda} \geq 0} L_P(\mathbf{x}^*, \boldsymbol{\lambda}) \quad \text{and} \quad L_D(\boldsymbol{\lambda}^*) \;=\; \inf_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}^*) \tag{A.133}$$

that the converse is true: if $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a saddle point of $L_p(\mathbf{x}, \boldsymbol{\lambda})$ then strong duality holds.

An optimal point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, under strong duality, simultaneously minimizes $L_P(\mathbf{x}, \boldsymbol{\lambda})$ with respect to $\mathbf{x}$ and maximizes $L_P(\mathbf{x}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$. In particular, an optimal point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfies

$$\mathbf{x}^* \;=\; \arg\min_{\mathbf{x} \in R^d} L_P(\mathbf{x}, \boldsymbol{\lambda}^*)\,. \tag{A.134}$$

Since this is an *unconstrained* minimization problem, necessary conditions for unconstrained minima apply. In particular, assuming that f and g_i are differentiable, for $i = 1, \ldots, n$, the general stationarity condition must be satisfied:

$$\nabla_{\mathbf{x}} L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;=\; \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) \;=\; 0\,. \tag{A.135}$$

Another consequence of (A.129) is

$$f(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* g_i(\mathbf{x}^*) \;\Rightarrow\; \sum_{i=1}^{n} \lambda_i^* g_i(\mathbf{x}^*) = 0\,, \tag{A.136}$$

from which the following important *complementary slackness* conditions follow:

$$\lambda_i^* g_i(\mathbf{x}^*) \;=\; 0\,, \quad i = 1, \ldots, n\,. \tag{A.137}$$

This means that if a constraint is inactive at the optimum, i.e., $g_i(\mathbf{x}^*) < 0$, then the corresponding optimal Lagrange multiplier λ_i^* must be zero. Conversely, $\lambda_i^* > 0$ implies that $g_i(\mathbf{x}^*) = 0$, i.e., the corresponding constraint is active (tight) at the optimum.

We can summarize all the previous results in the following classical theorem.

Theorem A.19. *(Karush-Kuhn-Tucker Conditions). Let $\mathbf{x}^*$ be a solution of the original optimization problem in (A.120), and let $\boldsymbol{\lambda}^*$ be a solution of the dual optimization problem in (A.126) such that strong duality is satisfied. Assume further that f and g_i are differentiable, for $i = 1, \ldots, n$. Then the following conditions must be satisfied:*

$$\begin{aligned}
&\nabla_{\mathbf{x}} L_P(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;=\; \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) \;=\; 0\,, && \text{(stationarity)} \\
&g_i(\mathbf{x}^*) \leq 0\,, \quad i = 1, \ldots, n\,, && \text{(primal feasibility)} \\
&\lambda_i^* \geq 0\,, \quad i = 1, \ldots, n\,, && \text{(dual feasibility)} \\
&\lambda_i^* g_i(\mathbf{x}^*) \;=\; 0\,, \quad i = 1, \ldots, n\,. && \text{(complementary slackness)}
\end{aligned} \tag{A.138}$$

Furthermore, it can be shown that if the original optimization problem in (A.120) is convex with affine constraints, then the KKT conditions are also sufficient for optimality.

A4 Proof of the Cover-Hart Theorem

In this section we present proofs of Thm 5.1 and 5.3. The proof of Thm 5.1 follows the general structure of the original proof in Cover and Hart [1967], with some differences. This proof assumes existence and continuity almost everywhere of the class-conditional densities. In Stone [1977] a more general proof is given, which does not assume existence of densities (see also Chapter 5 of Devroye et al. [1996]).

Proof of Theorem 5.1
First, one has to show that the nearest neighbor $\mathbf{X}_n^{(1)}$ of a test point $\mathbf{X}$ converges to $\mathbf{X}$ as $n \to \infty$. The existence of densities makes this simple to show. First note that, for any $\tau > 0$,

$$P(||\mathbf{X}_n^{(1)} - \mathbf{X}|| > \tau) \;=\; P(||\mathbf{X}_i - \mathbf{X}|| > \tau;\, i = 1, \ldots, n) \;=\; (1 - P(||\mathbf{X}_1 - \mathbf{X}|| < \tau))^n \,. \qquad \text{(A.139)}$$

If we can show that $P(||\mathbf{X}_1 - \mathbf{X}|| < \tau) > 0$, then it follows from (A.139) that $P(||\mathbf{X}_n^{(1)} - \mathbf{X}|| > \tau) \to 0$, so that $\mathbf{X}_n^{(1)} \to \mathbf{X}$ in probability. Since $\mathbf{X}_1$ and $\mathbf{X}$ are independent and identically distributed with density $p_{\mathbf{X}}$, $\mathbf{X}_1 - \mathbf{X}$ has a density $p_{\mathbf{X}_1 - \mathbf{X}}$, given by the classical convolution formula:

$$p_{\mathbf{X}_1 - \mathbf{X}}(\mathbf{x}) \;=\; \int_{R^d} p_{\mathbf{X}}(\mathbf{x} + \mathbf{u})\, p_{\mathbf{X}}(\mathbf{u})\, d\mathbf{u} \,. \qquad \text{(A.140)}$$

From this, we have $p_{\mathbf{X}_1 - \mathbf{X}}(\mathbf{0}) = \int_{R^d} p_{\mathbf{X}}^2(\mathbf{x})\, d\mathbf{u} > 0$. It follows, by continuity of the integral, that $p_{\mathbf{X}_1 - \mathbf{X}}$ must be nonzero in a neighborhood of $\mathbf{0}$, i.e., $P(||\mathbf{X}_1 - \mathbf{X}|| < \tau) > 0$, as was to be shown.

Now, let Y_n' denote the label of the nearest neighbor $\mathbf{X}_n^{(1)}$. Consider the conditional error rate

$$\begin{aligned} P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \;&=\; P(Y_n' \neq Y \mid \mathbf{X}, \mathbf{X}_n^{(1)}) \\ &=\; P(Y = 1, Y_n' = 0 \mid \mathbf{X}, \mathbf{X}_n^{(1)}) + P(Y = 0, Y_n' = 1 \mid \mathbf{X}, \mathbf{X}_n^{(1)}) \\ &=\; P(Y = 1 \mid \mathbf{X}) P(Y_n' = 0 \mid \mathbf{X}_n^{(1)}) + P(Y = 0 \mid \mathbf{X}) P(Y_n' = 1 \mid \mathbf{X}_n^{(1)}) \\ &=\; \eta(\mathbf{X})(1 - \eta(\mathbf{X}_n^{(1)})) + (1 - \eta(\mathbf{X}))\eta(\mathbf{X}_n^{(1)}) \end{aligned} \qquad \text{(A.141)}$$

where independence of $(\mathbf{X}_n^{(1)}, Y_n')$ and $(\mathbf{X}, Y)$ was used. We now use the assumption that the class-conditional densities exist and are continuous a.e., which implies that η is continuous a.e. We had established previously that $\mathbf{X}_n^{(1)} \to \mathbf{X}$ in probability. By the Continuous Mapping Theorem (see Theorem A.6), $\eta(\mathbf{X}_n^{(1)}) \to \eta(\mathbf{X})$ in probability and

$$P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \;\to\; 2\eta(\mathbf{X})(1 - \eta(\mathbf{X})) \;\text{ in probability.} \qquad \text{(A.142)}$$

Since all random variables are bounded in the interval $[0, 1]$, we can apply the Bounded Convergence Theorem (see Thm. A.11) to obtain

$$E[\varepsilon_n] \;=\; E[P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n)] \;\to\; E[2\eta(\mathbf{X})(1 - \eta(\mathbf{X}))] \,, \qquad \text{(A.143)}$$

proving the first part of the theorem.

For the second part, let $r(\mathbf{X}) = \min\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\}$ and note that $\eta(\mathbf{X})(1-\eta(\mathbf{X})) = r(\mathbf{X})(1-r(\mathbf{X}))$. It follows that

$$\begin{aligned} \varepsilon_{\mathrm{NN}} &= E[2\eta(\mathbf{X})(1-\eta(\mathbf{X}))] = E[2r(\mathbf{X})(1-r(\mathbf{X}))] \\ &= 2E[r(\mathbf{X})]E[(1-r(\mathbf{X}))] + 2\mathrm{Cov}(r(\mathbf{X}), 1-r(\mathbf{X})) \\ &= 2\varepsilon^*(1-\varepsilon^*) - 2\mathrm{Var}(r(\mathbf{X})) \leq 2\varepsilon^*(1-\varepsilon^*) \leq 2\varepsilon^*, \end{aligned} \tag{A.144}$$

as required.

Proof of Theorem 5.3

The proof of (5.13) and (5.14) follows the same structure as in the case $k = 1$. As before, the first step is to show that the ith-nearest neighbor $\mathbf{X}_n^{(i)}$ of $\mathbf{X}$, for $i = 1, \ldots, k$, converges to $\mathbf{X}$ in probability as $n \to \infty$. This is so because, for every $\tau > 0$,

$$P(||\mathbf{X}_n^{(i)} - \mathbf{X}|| > \tau) = P(||\mathbf{X}_j - \mathbf{X}|| > \tau;\, j = k, \ldots, n) = (1 - P(||\mathbf{X}_1 - \mathbf{X}|| < \tau))^{n-k-1} \to 0\,, \tag{A.145}$$

since $P(||\mathbf{X}_1 - \mathbf{X}|| < \tau) > 0$, as shown in the previous proof. Next, let the label of the $i$the-nearest neighbor $\mathbf{X}_n^{(i)}$ of $\mathbf{X}$ by $Y_n^{(i)}$, and consider the conditional error rate

$$\begin{aligned} &P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \\ &= P(Y = 1, \textstyle\sum_{i=1}^k Y_n^{(i)} < \frac{k}{2} \mid \mathbf{X}, \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) + P(Y = 0, \textstyle\sum_{i=1}^k Y_n^{(i)} > \frac{k}{2} \mid \mathbf{X}, \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) \\ &= P(Y = 1 \mid \mathbf{X}) P(\textstyle\sum_{i=1}^k Y_n^{(i)} < \frac{k}{2} \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) \\ &\quad + P(Y = 0 \mid \mathbf{X}) P(\textstyle\sum_{i=1}^k Y_n^{(i)} > \frac{k}{2} \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) \\ &= \eta(\mathbf{X}) \textstyle\sum_{i=0}^{(k-1)/2} P(\sum_{j=1}^k Y_n^{(j)} = i \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) \\ &\quad + (1 - \eta(\mathbf{X})) \textstyle\sum_{i=(k+1)/2}^k P(\sum_{j=1}^k Y_n^{(j)} = i \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)})\,, \end{aligned} \tag{A.146}$$

where

$$\begin{aligned} P(\textstyle\sum_{j=1}^k Y_n^{(j)} = i \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) &= \sum_{\substack{m_1, \ldots, m_k \in \{0,1\} \\ m_1 + \cdots + m_k = i}} \prod_{j=1}^k P(Y_n^{(j)} = m_j \mid \mathbf{X}_n^{(j)}) \\ &= \sum_{\substack{m_1, \ldots, m_k \in \{0,1\} \\ m_1 + \cdots + m_k = i}} \prod_{j=1}^k \eta(\mathbf{X}_n^{(j)})^{m_j} (1 - \eta(\mathbf{X}_n^{(j)}))^{1-m_j}. \end{aligned} \tag{A.147}$$

Using the previously established fact that $\mathbf{X}_n^{(j)} \to \mathbf{X}$ in probability, for $i = 1, \ldots, k$, it follows from the assumption of continuity of the distributions a.e. and the Continuous Mapping Theorem

(see Theorem A.6) that

$$P(\textstyle\sum_{j=1}^{k} Y_n^{(j)} = i \mid \mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(k)}) \xrightarrow{P} \sum_{\substack{m_1,\ldots,m_k \in \{0,1\} \\ m_1+\cdots+m_k=i}} \prod_{j=1}^{k} \eta(\mathbf{X})^{m_j}(1-\eta(\mathbf{X}))^{1-m_j} = \binom{k}{i} \eta(\mathbf{X})^i (1-\eta(\mathbf{X}))^{k-i} \tag{A.148}$$

and

$$P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \xrightarrow{P} \textstyle\sum_{i=0}^{(k-1)/2} \eta(\mathbf{X})^{i+1}(1-\eta(\mathbf{X}))^{k-i} + \sum_{i=(k+1)/2}^{k} \eta(\mathbf{X})^i (1-\eta(\mathbf{X}))^{k+1-i} . \tag{A.149}$$

Since all random variables are bounded in the interval $[0, 1]$, we can apply the Bounded Convergence Theorem (see Thm. A.11) to obtain

$$\begin{aligned} E[\varepsilon_n] &= E[P(\psi_n(\mathbf{X}) \neq Y \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n)] \\ &\to E\left[\textstyle\sum_{i=0}^{(k-1)/2} \eta(\mathbf{X})^{i+1}(1-\eta(\mathbf{X}))^{k-i} + \sum_{i=(k+1)/2}^{k} \eta(\mathbf{X})^i (1-\eta(\mathbf{X}))^{k+1-i}\right], \end{aligned} \tag{A.150}$$

establishing (5.13) and (5.14).

For the second part, as before, we let $r(\mathbf{X}) = \min\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\}$ and note that $\eta(\mathbf{X})(1-\eta(\mathbf{X})) = r(\mathbf{X})(1-r(\mathbf{X}))$. By symmetry, it is easy to see that $\alpha_k(\eta(\mathbf{X})) = \alpha_k(r(\mathbf{X}))$. We seek an inequality $\alpha_k(r(\mathbf{X})) \leq a_k r(\mathbf{X})$, so that

$$\varepsilon_{\text{kNN}} = E[\alpha_k(\eta(\mathbf{X}))] = E[\alpha_k(r(\mathbf{X}))] \leq a_k E[r(\mathbf{X})] = a_k \varepsilon^* , \tag{A.151}$$

where $a_k > 1$ is as small as possible. But as can be seen in Figure 5.8, a_k corresponds to the slope of the tangent line to $\alpha_k(p)$, in the range $p \in [0, \frac{1}{2}]$, through the origin, so it must satisfy (5.21).

A5 Proof of Stone's Theorem

In this section, we present a proof of Thm 5.4, which essentially follows the proof given by Devroye et al. [1996]. The original proof in Stone [1977] is more general, relaxing the nonnegativity and normalization assumptions (5.2) on the weights, while also showing that, under (5.2), the conditions on the weights given in the theorem are both necessary and sufficient for universal consistency.

Proof of Theorem 5.4

It follows from Lemma 5.1, and the comment following it, that it is sufficient to show that $E|(\eta_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \to 0$, as $n \to \infty$. Introduce the smoothed posterior-probability function

$$\tilde{\eta}_n(\mathbf{x}) = \sum_{i=1}^{n} W_{n,i}(\mathbf{x}) \eta(\mathbf{X}_i) . \tag{A.152}$$

This is not a true estimator, since it is a function of $\eta(\mathbf{x})$. However, it allows one to break the problem down into two manageable parts:

$$\begin{aligned} E[(\eta_n(\mathbf{X}) - \eta(\mathbf{X}))^2] &= E[(\eta_n(\mathbf{X}) - \tilde{\eta}_n(\mathbf{X}) + \tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \\ &\leq 2E[(\eta_n(\mathbf{X}) - \tilde{\eta}_n(\mathbf{X}))^2] + 2E[(\tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2], \end{aligned} \tag{A.153}$$

where the inequality follows from the fact that $(a+b)^2 \leq 2(a^2+b^2)$. The rest of the proof consists in showing that $E[(\eta_n(\mathbf{X}) - \tilde{\eta}_n(\mathbf{X}))^2] \to 0$, and then showing that $E[(\tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \to 0$.

For the first part, notice that

$$\begin{aligned} E[(\eta_n(\mathbf{X}) - \tilde{\eta}_n(\mathbf{X}))^2] &= E\left[\left(\sum_{i=1}^n W_{ni}(\mathbf{X})(Y_i - \eta(\mathbf{X}_i)\right)^2\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n E\left[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - \eta(\mathbf{X}_i)(Y_j - \eta(\mathbf{X}_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n E\left[E\left[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - \eta(\mathbf{X}_i)(Y_j - \eta(\mathbf{X}_j) \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\right]\right] \end{aligned} \tag{A.154}$$

Now, given $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$, $W_{ni}(\mathbf{X})$ and $W_{nj}(\mathbf{X})$ are constants, and $Y_i - \eta(\mathbf{X}_i)$ and $Y_j - \eta(\mathbf{X}_j)$ are zero-mean random variables. Furthermore, $Y_i - \eta(\mathbf{X}_i)$ and $Y_j - \eta(\mathbf{X}_j)$ are independent if $i \neq j$. Therefore, $E\left[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - \eta(\mathbf{X}_i)(Y_j - \eta(\mathbf{X}_j) \mid \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = 0$, for $i \neq j$, and we obtain

$$\begin{aligned} E[(\eta_n(\mathbf{X}) - \tilde{\eta}_n(\mathbf{X}))^2] &= \sum_{i=1}^n E\left[W_{ni}^2(\mathbf{X})(Y_i - \eta(\mathbf{X}_i)^2\right] \\ &\leq E\left[\sum_{i=1}^n W_{ni}^2(\mathbf{X})\right] \leq E\left[\max_{i=1,\ldots,n} W_{n,i}(\mathbf{x}) \sum_{i=1}^n W_{ni}(\mathbf{X})\right] = E\left[\max_{i=1,\ldots,n} W_{n,i}(\mathbf{x})\right] \to 0, \end{aligned} \tag{A.155}$$

by condition (ii) of Stone's Theorem and the Bounded Convergence Theorem A.11.

The second part is more technical. First, given $\tau > 0$, find a function η^* such that $0 \leq \eta^*(\mathbf{x}) \leq 1$, η^* is $P_{\mathbf{X}}$-square-integrable, continuous, and has compact support, and $E[(\eta^*(\mathbf{X}) - \eta(\mathbf{X}))^2] < \tau$. Such a function exists, because $\eta(\mathbf{x})$ is $P_{\mathbf{X}}$-integrable (see Section 2.6.3), and therefore square-integrable, since $\eta^2(\mathbf{x}) \leq \eta(\mathbf{x})$, and the set of continuous function with compact support is dense in the set of

square-integrable functions. Now, write

$$
\begin{aligned}
E[(\tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \;&=\; E\left[\left(\sum_{i=1}^n W_{ni}(\mathbf{X})(\eta(\mathbf{X}_i) - \eta(\mathbf{X}))\right)^2\right] \;\leq\; E\left[\sum_{i=1}^n W_{ni}(\mathbf{X})(\eta(\mathbf{X}_i) - \eta(\mathbf{X}))^2\right] \\
&= E\left[\sum_{i=1}^n W_{ni}(\mathbf{X})\left((\eta(\mathbf{X}_i) - \eta^*(\mathbf{X}_i)) + (\eta^*(\mathbf{X}_i) - \eta^*(\mathbf{X})) + (\eta^*(\mathbf{X}) - \eta(\mathbf{X}))\right)^2\right] \\
&\leq 3E\left[\sum_{i=1}^n W_{ni}(\mathbf{X})\left((\eta(\mathbf{X}_i) - \eta^*(\mathbf{X}_i))^2 + (\eta^*(\mathbf{X}_i) - \eta^*(\mathbf{X}))^2 + (\eta^*(\mathbf{X}) - \eta(\mathbf{X}))^2\right)\right] \\
&\leq 3E\left[\sum_{i=1}^n W_{ni}(\mathbf{X})(\eta(\mathbf{X}_i) - \eta^*(\mathbf{X}_i))^2\right] + 3E\left[\sum_{i=1}^n W_{ni}(\mathbf{X})(\eta^*(\mathbf{X}_i) - \eta^*(\mathbf{X}))^2\right] + 3E\left[(\eta^*(\mathbf{X}) - \eta(\mathbf{X}))^2\right] \\
&= I + II + III\,,
\end{aligned}
\tag{A.156}
$$

where the first inequality follows from Jensen's Inequality, while the second inequality follows from the fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$. Now, by construction of η^* and condition (iii) of Stone's Theorem, it follows that $I < 3\tau$ and $III < 3c\tau$. To bound II, notice that η^*, being continuous on a compact support, is also uniformly continuous. Hence, given $\tau > 0$, there is a $\delta > 0$ such that $||\mathbf{x}' - \mathbf{x}|| < \delta$ implies that $|\eta^*(\mathbf{x}') - \eta^*(\mathbf{x})| < \tau$, for all $\mathbf{x}', \mathbf{x} \in R^d$. Hence,

$$
II \;\leq\; 3E\left[\sum_{i=1}^n W_{n,i}(\mathbf{X}) I_{||\mathbf{X}_i - \mathbf{X}|| > \delta}\right] + 3E\left[\sum_{i=1}^n W_{n,i}(\mathbf{X})\tau\right] \;=\; 3E\left[\sum_{i=1}^n W_{n,i}(\mathbf{X}) I_{||\mathbf{X}_i - \mathbf{X}|| > \delta}\right] + 3\tau\,,
\tag{A.157}
$$

where we used the fact that $|\eta^*(\mathbf{x}') - \eta^*(\mathbf{x})| \leq 1$. Using condition (i) of Stone's Theorem and the Bounded Convergence Theorem A.11, it follows that $\limsup_{n\to\infty} II \leq 3\tau$. Putting all together,

$$
\limsup_{n\to\infty} E[(\tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \leq 3\tau + 3c\tau + 3\tau = 3(c+2)\tau\,.
\tag{A.158}
$$

Since τ is arbitrary, it follows that $E[(\tilde{\eta}_n(\mathbf{X}) - \eta(\mathbf{X}))^2] \to 0$ and the proof is complete.

A6 Proof of the Vapnik-Chervonenkis Theorem

In this section, we present a proof of Thm 8.2. Our proof combines elements of the proofs given by Pollard [1984] and Devroye et al. [1996], who credit Dudley [1978]. See also Castro [2020]. We prove a general version of the result and then specialize it to the classification case.

Consider a probability space $(R^p, \mathcal{B}^p, \nu)$, and n i.i.d. random variables $Z_1, \ldots, Z_n \sim \nu$. (For a review of probability theory, see Section A1.) Note that each Z_i is in fact a random vector, but we do not employ the usual boldface type here, so as not to encumber the notation. An *empirical measure* is

a random measure on $(R^p, \mathcal{B}^P)$ that is a function of $Z_1, \ldots, Z_n$. The standard empirical measure ν_n puts mass $1/n$ over each Z_i, so that

$$\nu_n(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{Z_i \in A}\,, \tag{A.159}$$

for $A \in \mathcal{B}^p$. By the Law of Large Numbers (LLN), $\nu_n(A) \stackrel{a.s.}{\longrightarrow} \nu(A)$, as $n \to \infty$, for any fixed A. In the VC theorem, one is interested instead in a *uniform* version of the LLN: $\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \stackrel{a.s.}{\longrightarrow} 0$, for a suitably provided family of sets $\mathcal{A} \subset \mathcal{B}^p$. General conditions to ensure the measurability of $\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|$ and of various other quantities in the proofs are discussed in Pollard [1984]; such will be assumed tacitly below.

Define a second (signed) empirical measure $\tilde{\nu}_n$, which puts mass $1/n$ or $-1/n$ randomly over each Z_i, i.e.,

$$\tilde{\nu}_n(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} \sigma_i I_{Z_i \in A} \tag{A.160}$$

for $A \in \mathcal{A}$, where $\sigma_1, \ldots, \sigma_n$ are i.i.d. random variables with $P(\sigma_1 = 1) = P(\sigma_1 = -1) = 1/2$, independently of $Z_1, \ldots, Z_n$.

It turns out that the VC theorem, much as Theorem 8.1, can be proved by a direct application of the Union Bound (A.10) and Hoeffding's Inequality (8.8), with the addition of the next key lemma.

Lemma A.2. *(Symmetrization Lemma). Regardless of the measure ν,*

$$P\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau\right) \le 4P\left(\sup_{A\in\mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4}\right), \quad \text{for all } \tau > 0 \text{ and } n \ge 2\tau^{-2}\,. \tag{A.161}$$

Proof. Consider a second sample $Z'_1, \ldots, Z'_n \sim \nu$, independent of $Z_1, \ldots, Z_n$ and the signs $\sigma_1, \ldots, \sigma_n$. In the first part of the proof, one seeks to relate the tail probability of $\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|$ in (A.161) to a tail probability of $\sup_{A\in\mathcal{A}} |\nu'_n(A) - \nu_n(A)|$, where

$$\nu'_n(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{Z'_i \in A}\,, \tag{A.162}$$

for $A \in \mathcal{A}$, and, in the second part, relate that to the tail probability of $\sup_{A\in\mathcal{A}} |\tilde{\nu}_n(A)|$ in (A.161).

Notice that, whenever $\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau$, there is an $A^* \in \mathcal{A}$, which is a function of $Z_1, \ldots, Z_n$, such that $|\nu_n(A^*) - \nu(A^*)| > \tau$, with probability 1. In other words,

$$P\left(|\nu_n(A^*) - \nu(A^*)| > \tau \;\middle|\; \sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau\right) \;=\; 1\,, \tag{A.163}$$

which in turn implies that

$$P\left(|\nu_n(A^*) - \nu(A^*)| > \tau\right) \;\ge\; P\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau\right). \tag{A.164}$$

Now, conditioned on $Z_1, \ldots, Z_n$, $\mathcal{A}^*$ is fixed (nonrandom). Notice that $E[\nu'_n(A^*) \mid Z_1, \ldots, Z_n] = \nu(A^*)$ and $\mathrm{Var}(\nu'_n(A^*) \mid Z_1, \ldots, Z_n) = \nu(A^*)(1 - \nu(A*))/n$. Hence, we can apply Chebyshev's Inequality (A.75) to get:

$$P\left(|\nu'_n(A^*) - \nu(A^*)| < \frac{\tau}{2} \,\middle|\, Z_1, \ldots, Z_n\right) \geq 1 - \frac{4\nu(A^*)(1-\nu(A^*))}{n\tau^2} \geq 1 - \frac{1}{n\tau^2} \geq \frac{1}{2}, \tag{A.165}$$

for $n \geq 2\tau^{-2}$. Now,

$$\begin{aligned} P\left(\sup_{A\in\mathcal{A}} |\nu'_n(A) - \nu_n(A)| > \frac{\tau}{2} \,\middle|\, Z_1, \ldots, Z_n\right) &\geq P\left(|\nu'_n(A^*) - \nu_n(A^*)| > \frac{\tau}{2} \,\middle|\, Z_1, \ldots, Z_n\right) \\ \geq I_{|\nu_n(A^*)-\nu(A^*)|>\tau}\, P\left(|\nu'_n(A^*) - \nu(A^*)| < \frac{\tau}{2} \,\middle|\, Z_1, \ldots, Z_n\right) &\geq \frac{1}{2} I_{|\nu_n(A^*)-\nu(A^*)|>\tau}\,. \end{aligned} \tag{A.166}$$

where the second inequality follows from the fact that $|a - c| > \tau$ and $|b - c| < \tau/2$ imply that $|a - b| > \tau/2$. Integrating (A.166) on both sides with respect to $Z_1, \ldots, Z_n$ and using (A.164) yields

$$P\left(\sup_{A\in\mathcal{A}} |\nu'_n(A) - \nu_n(A)| > \frac{\tau}{2}\right) \geq \frac{1}{2} P\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau\right), \tag{A.167}$$

which completes the first part of the proof. Next, define

$$\tilde{\nu}'_n(A) = \frac{1}{n}\sum_{i=1}^{n} \sigma_i I_{Z'_i \in A} \tag{A.168}$$

for $A \in \mathcal{A}$. The key observation at this point is that $\sup_{A\in\mathcal{A}} |\nu'_n(A) - \nu_n(A)|$ has the same distribution as $\sup_{A\in\mathcal{A}} |\tilde{\nu}'_n(A) - \tilde{\nu}_n(A)|$, which can be seen by conditioning on $\sigma_1, \ldots, \sigma_n$. Hence,

$$\begin{aligned} P\left(\sup_{A\in\mathcal{A}} |\nu'_n(A) - \nu_n(A)| > \frac{\tau}{2}\right) &= P\left(\sup_{A\in\mathcal{A}} |\tilde{\nu}'_n(A) - \tilde{\nu}_n(A)| > \frac{\tau}{2}\right) \\ &\leq P\left(\left\{\sup_{A\in\mathcal{A}} |\tilde{\nu}'_n(A)| > \frac{\tau}{4}\right\} \bigcup \left\{\sup_{A\in\mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4}\right\}\right) \\ &\leq P\left(\sup_{A\in\mathcal{A}} |\tilde{\nu}'_n(A)| > \frac{\tau}{4}\right) + P\left(\sup_{A\in\mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4}\right) = 2P\left(\sup_{A\in\mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4}\right), \end{aligned} \tag{A.169}$$

where the first inequality follows from the fact that $|a - b| > \tau/2$ implies that $|a| > \tau/4$ or $|b| > \tau/4$, while the second inequality is an application of the Union Bound (A.10). Combining (A.167) and (A.169) proves the lemma. ⋄

Equipped with the Symmetrization Lemma, the proof of the following theorem is fairly simple, but also quite instructive.

Theorem A.20. *(General Vapnik-Chervonenkis Theorem.) Regardless of the measure ν,*

$$P\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \tau\right) \leq 8\,\mathcal{S}(\mathcal{A}, n) e^{-n\tau^2/32}, \quad \text{for all } \tau > 0\,. \tag{A.170}$$

where $\mathcal{S}(\mathcal{A}, n)$ is the nth shatter coefficient of $\mathcal{A}$, defined in (8.14).

Proof. For fixed $Z_1 = z_1, \ldots, Z_n = z_n$, consider the binary vector $(I_{z_i \in A}, \ldots, I_{z_i \in A})$, as A ranges over $\mathcal{A}$. There are of course a maximum of 2^n distinct values that this vector can take on. But, for a given $\mathcal{A}$, this number may be smaller than 2^n. Indeed, this is the number $N_{\mathcal{A}}(z_1, \ldots, z_n)$, defined in (8.13) — by definition, this number must be smaller than the shatter coefficient $\mathcal{S}(\mathcal{A}, n)$, for any choice of $z_1, \ldots, z_n$. Notice that $\tilde{\nu}_n(A)$, conditioned on $Z_1 = z_1, \ldots, Z_n = z_n$, is still a random variable, through the random signs $\sigma_1, \ldots, \sigma_n$. Since this random variable is a function of the vector $(I_{z_i \in A}, \ldots, I_{z_i \in A})$, the number of values it can take as A ranges over $\mathcal{A}$ is also bounded by $\mathcal{S}(\mathcal{A}, n)$. Therefore, $\sup_{A \in \mathcal{A}} |\tilde{\nu}_n(A)|$ turns out to be a maximum of at most $\mathcal{S}(\mathcal{A}, n)$ values, so that one can employ the Union Bound (A.10) as follows:

$$\begin{aligned} P\left(\sup_{A \in \mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4} \,\middle|\, Z_1, \ldots, Z_n\right) &= P\left(\bigcup_{A \in \mathcal{A}} \left\{|\tilde{\nu}_n(A)| > \frac{\tau}{4}\right\} \,\middle|\, Z_1, \ldots, Z_n\right) \\ \leq \sum_{A \in \mathcal{A}} P\left(|\tilde{\nu}_n(A)| > \frac{\tau}{4} \,\middle|\, Z_1, \ldots, Z_n\right) &\leq \mathcal{S}(\mathcal{A}, n) \sup_{A \in \mathcal{A}} P\left(|\tilde{\nu}_n(A)| > \frac{\tau}{4} \,\middle|\, Z_1, \ldots, Z_n\right), \end{aligned} \tag{A.171}$$

with the understanding that the union, sum, and suprema are finite. Now we apply Hoeffding's Inequality (Theorem A.14) to bound the probability $P\left(|\tilde{\nu}_n(A)| > \frac{\tau}{4} \,\middle|\, Z_1, \ldots, Z_n\right)$. Conditioned on $Z_1 = z_1, \ldots, Z_n = z_n$, $\tilde{\nu}_n(A) = \sum_{i=1}^n \sigma_i I_A(z_i \in A)$ is a sum of independent zero-mean random variables, which are bounded in the interval $[-1, 1]$ (they are not identically-distributed, but this is not necessary for application of Theorem A.14). Hoeffding's Inequality then yields:

$$P\left(|\tilde{\nu}_n(A)| > \frac{\tau}{4} \,\middle|\, Z_1, \ldots, Z_n\right) \leq 2e^{-n\tau^2/32}, \quad \text{for all } \tau > 0\,. \tag{A.172}$$

Applying (A.171) and integrating on both sides with respect to $Z_1, \ldots, Z_n$ yields

$$P\left(\sup_{A \in \mathcal{A}} |\tilde{\nu}_n(A)| > \frac{\tau}{4}\right) \leq 2\,\mathcal{S}(\mathcal{A}, n) e^{-n\tau^2/32}, \quad \text{for all } \tau > 0\,. \tag{A.173}$$

Now, if $n < 2\tau^{-2}$, the inequality in (A.170) is trivial. If $n \geq 2\tau^{-2}$, we can apply Lemma A.2 and get the desired result. $\diamond$

If $\mathcal{S}(\mathcal{A}, n)$ grows polynomially with n (this is the case if the VC dimension of $\mathcal{A}$ is finite), then, by an application of Theorem A.8, (A.170) yields the uniform LLN: $\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| \xrightarrow{a.s.} 0$.

Specializing Theorem A.20 to the classification case yields the required proof.

Proof of Theorem 8.2

Consider the probability space $(R^{d+1}, \mathcal{B}^{d+1}, P_{\mathbf{X},Y})$, where $P_{\mathbf{X},Y}$ is the joint feature-label probability measure constructed in Section 2.6.3. Let the i.i.d. training data be $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. Given a family of classifiers $\mathcal{C}$, apply Theorem A.20 with $\nu = P_{\mathbf{X},Y}$, $Z_i = (\mathbf{X}_i, Y_i) \sim P_{\mathbf{X},Y}$, for $i = 1, \ldots, n$, and $\tilde{\mathcal{A}}_{\mathcal{C}}$ containing all set of the kind

$$\tilde{A}_\psi = \{\psi(\mathbf{X}) \neq Y\} = \{\psi(\mathbf{X}) = 1, Y = 0\} \cup \{\psi(\mathbf{X}) = 0, Y = 1\}, \tag{A.174}$$

for each $\psi \in \mathcal{C}$ (the sets $\tilde{A}_\psi$ are Borel since classifiers are measurable functions). Then $\nu(\tilde{A}_\psi) = \varepsilon[\psi]$, $\nu_n(\tilde{A}_\psi) = \hat{\varepsilon}[\psi]$, and $\sup_{\tilde{A}_\psi \in \tilde{\mathcal{A}}_\mathcal{C}} |\nu_n(\tilde{A}_\psi) - \nu(\tilde{A}_\psi)| = \sup_{\psi \in \mathcal{C}} |\hat{\varepsilon}[\psi] - \varepsilon[\psi]|$. It remains to show that $\mathcal{S}(\tilde{\mathcal{A}}_\mathcal{C}, n) = \mathcal{S}(\mathcal{A}_\mathcal{C}, n)$, where $\mathcal{A}_\mathcal{C} = \{A_\psi \mid \psi \in \mathcal{C}\}$, and A_ψ is defined in (8.23). First note that there is a one-to-one correspondence between $\tilde{\mathcal{A}}_\mathcal{C}$ and $\mathcal{A}_\mathcal{C}$, since, for each $\psi \in \mathcal{C}$, we have $\tilde{A}_\psi = A_\psi \times \{0\} \cup A_\psi^c \times \{1\}$. Given a set of points $\{x_1, \ldots, x_n\}$, if k points are picked by A_ψ, then k points can be picked by $\tilde{A}_\psi$ in the set $\{(x_1, 1), \ldots, (x_n, 1)\}$; hence $\mathcal{S}(\mathcal{A}_\mathcal{C}, n) \leq \mathcal{S}(\tilde{\mathcal{A}}_\mathcal{C}, n)$. On the other hand, given a set of points $\{(x_1, 0), \ldots, (x_{n_0}, 0), (x_{n_0+1}, 1), \ldots, (x_{n_0+n_1}, 1)\}$, suppose that $\tilde{A}_\psi$ picks out the subset $\{(x_1, 0), \ldots, (x_l, 0), (x_{n_0+1}, 1), \ldots, (x_{n_0+m}, 1)\}$ (the sets can be unambiguously written this way, since order does not matter). Then A_ψ picks out the subset $\{(x_1, \ldots, x_l, x_{n_0+m+1}, x_{n_0+n_1}\}$, among the set of points $\{x_1, \ldots, x_{n_0+n_1}\}$, and the two subsets determine each other uniquely, so $\mathcal{S}(\tilde{\mathcal{A}}_\mathcal{C}, n) \leq \mathcal{S}(\mathcal{A}_\mathcal{C}, n)$ Thus, $\mathcal{S}(\tilde{\mathcal{A}}_\mathcal{C}, n) = \mathcal{S}(\mathcal{A}_\mathcal{C}, n)$. (Thus, the VC dimensions also agree: $V_{\tilde{\mathcal{A}}_\mathcal{C}} = V_{\mathcal{A}_\mathcal{C}}$.)

A7 Proof of Convergence of the EM Algorithm

Here we present a proof of convergence of the general Expectation-Maximization algorithm to a local maximum of the log-likelihood function.

Let $\mathbf{X}$, $\mathbf{Z}$, $\boldsymbol{\theta} \in \Theta$be the observed data, the hidden variables, and the vector of model, respectively. meters. The EM method relies on a clever application of Jensen's inequality to obtain the following lower bound on the "incomplete" log-likelihood $L(\boldsymbol{\theta}) = \ln p_{\boldsymbol{\theta}}(\mathbf{X})$:

$$B(\boldsymbol{\theta}) \;=\; \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X})}{q(\mathbf{Z})} \;\leq\; \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X})}{q(\mathbf{Z})} \;=\; \ln \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X}) \;=\; L(\boldsymbol{\theta})\,, \tag{A.175}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\theta}$, where $q(\mathbf{Z})$ is an arbitrary probability distribution to be specified shortly. The inequality follows directly from concavity of the logarithm function and Jensen's inequality.

One would like to maximize the lower bound function $B(\boldsymbol{\theta})$ so that it touches $L(\boldsymbol{\theta})$, at a value $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$. We show by inspection that the choice $q(\mathbf{Z}; \boldsymbol{\theta}^{(m)}) = p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X})$ accomplishes this. First we replace this choice of $q(\mathbf{Z})$ in (A.175) to obtain:

$$B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \;=\; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln \frac{p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X})}{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X})}\,. \tag{A.176}$$

Now we verify that indeed this lower bound touches the log-likelihood at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$:

$$\begin{aligned} B(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \;&=\; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln \frac{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z}, \mathbf{X})}{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X})} \;=\; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln p_{\boldsymbol{\theta}^{(m)}}(\mathbf{X}) \\ &=\; \ln p_{\boldsymbol{\theta}^{(m)}}(\mathbf{X}) \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \;=\; L(\boldsymbol{\theta}^{(m)})\,. \end{aligned} \tag{A.177}$$

The main idea behind EM is that choosing a value of $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m+1)}$ that increases $B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ over its previous value $B(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$ will also increase $L(\boldsymbol{\theta})$ over its previous value $L(\boldsymbol{\theta}^{(m)})$. This can be proved as follows:

$$\begin{aligned} B(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) \; - \; B(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \; &= \; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln \frac{p_{\boldsymbol{\theta}^{(m+1)}}(\mathbf{Z}, \mathbf{X})}{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z}, \mathbf{X})} \\ &= \; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln \frac{p_{\boldsymbol{\theta}^{(m+1)}}(\mathbf{Z} \mid \mathbf{X})}{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X})} \; + \; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln \frac{p_{\boldsymbol{\theta}^{(n+1)}}(\mathbf{X})}{p_{\boldsymbol{\theta}^{(m)}}(\mathbf{X})} \\ &= \; - \, D(p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \,||\, p_{\boldsymbol{\theta}^{(m+1)}}(\mathbf{Z} \mid \mathbf{X})) \; + \; L(\boldsymbol{\theta}^{(m+1)}) - L(\boldsymbol{\theta}^{(m)}) \,, \end{aligned} \tag{A.178}$$

where $D(p \,||\, q)$ is the *Kullback-Leibler distance* between two probability mass functions. The KL distance is always nonnegative [Kullback, 1968], with equality if and only if $p = q$ with probability 1. We conclude that

$$B(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - B(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \leq \; L(\boldsymbol{\theta}^{(m+1)}) - L(\boldsymbol{\theta}^{(m)}) \,, \tag{A.179}$$

and that setting

$$\boldsymbol{\theta}^{(m+1)} \; = \; \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\theta}} B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \,, \tag{A.180}$$

will increase the log-likelihood $L(\boldsymbol{\theta})$, unless one is already at a local maximum of $L(\boldsymbol{\theta})$.[2] This fact is graphically represented in Figure A.4. This proves the eventual convergence of the EM procedure to a local maximum of $L(\boldsymbol{\theta})$. Now,

$$B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \; = \; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X}) \; - \; \sum_{\mathbf{Z}} p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \ln p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \,. \tag{A.181}$$

Since the second term in the previous equation does not depend on $\boldsymbol{\theta}$, the maximization in (A.180) can be accomplished by maximizing the first term only:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \; = \; \sum_{\mathbf{Z}} \ln p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X}) \, p_{\boldsymbol{\theta}^{(m)}}(\mathbf{Z} \mid \mathbf{X}) \; = \; E_{\boldsymbol{\theta}^{(m)}}[\ln p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{X}) \mid \mathbf{X}] \,. \tag{A.182}$$

The unknown hidden variable $\mathbf{Z}$ is "averaged out" by the expectation.

The resulting EM procedure consists of picking an initial guess $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ and iterating two steps:

- **E-Step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- **M-Step:** Find $\boldsymbol{\theta}^{(m+1)} \; = \; \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

for $n = 0, 1, \ldots$ until the improvement in the log-likelihood $|\ln L(\boldsymbol{\theta}^{(m+1)}) - \ln L(\boldsymbol{\theta}^{(m)})|$ falls below a pre-specified positive value.

[2] In fact, just selecting $\boldsymbol{\theta}^{(m+1)}$ such that $B(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - B(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) > 0$ will do — this is called "Generalized Expectation Maximization"

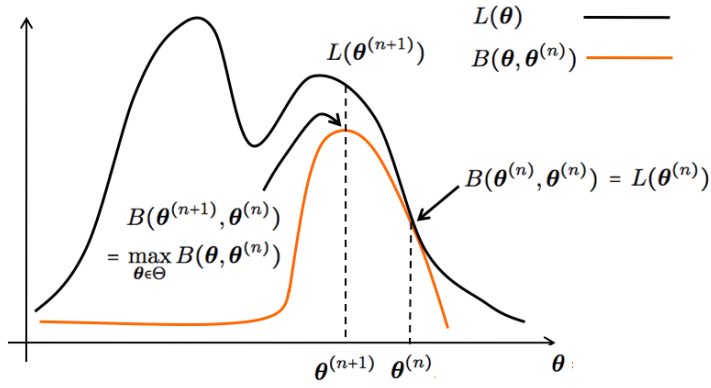

Figure A.4: The lower bound $B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ touches the log-likelihood $L(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$. Maximizing $B(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(m+1)}$ increases $L(\boldsymbol{\theta})$. Repeating the process leads to eventual convergence to a local maximum of $L(\boldsymbol{\theta})$. (Adapted from Figure 1 of Minka [1998].)

.

A8 Data Sets Used in the Book

In this section we describe the synthetic and real data sets that are used throughout the book. The real data sets can be downloaded from the book website.

A8.1 Synthetic Data

We employ a general multivariate Gaussian model to generate synthetic data, which consists of blocked covariance matrices of the form

$$\Sigma_{d\times d} = \begin{bmatrix} \Sigma_{l_1\times l_1} & 0 & \cdots & 0 \\ 0 & \Sigma_{l_2\times l_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{l_k\times l_k} \end{bmatrix} \tag{A.183}$$

where $l_1 + \cdots l_k = d$. The features are thus clustered into k independent groups. If $k = d$, then all features are independent. The individual covariance matrices$\Sigma_{\ l_i\times l_i}$ could be arbitrary, but here we will consider a simple parametric form

$$\Sigma_{l_i\times l_i}(\sigma_i^2, \rho_i) = \sigma_i^2 \begin{bmatrix} 1 & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \cdots & \rho_i \\ \vdots & \vdots & \ddots & \vdots \\ \rho_i & \rho_i & \cdots & 1 \end{bmatrix} \tag{A.184}$$

for $i = 1, \ldots, k$, where $-1 < \rho_i < 1$. Hence, the features within each block have the same variance σ_i^2 and are all correlated with the same correlation coefficient ρ_i.

The class mean vectors $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and prior probabilities $c_0 = P(Y = 0)$ and $c_1 = P(Y = 1)$ are arbitrary. Heteroskedastic Gaussian models result from specifying the class-conditional covariance matricesΣ_0 andΣ_1 separately. "Noisy features" can be obtained by matching mean components across classes and matching corresponding singleton blocks in the covariance matrices. Each noisy feature is an independent feature with the same mean and variance across the classes.

The python script `app_synth_data.py` generates sample data from this model.

A8.2 Dengue Fever Prognosis Data Set

This is gene-expression microarray data from a dengue fever diagnosis study performed in the Northeast of Brazil. The primary purpose of the study was to be able to predict the ultimate clinical outcome of dengue (whether the benign classical form or the dangerous hemorrhagic fever) from gene expression profiles of peripheral blood mononuclear cells (PBMCs) of patients in the early days of fever. The study is reported in Nascimento et al. [2009]. See also Example 1.1. The data consist of 26 training points measured on 1981 genes and three class labels, corresponding to: 8 classical dengue fever (DF) patients, 10 dengue hemorrhagic fever (DHF) patients, and 8 febrile non-dengue (ND) patients, as classified by an experienced clinician. This is a retrospective study, meaning that the patients were tracked and their outcomes verified by a clinician, but their status could not be determined clinically at the time the data was obtained, which was within one week of the start of symptoms.

A8.3 Breast Cancer Prognosis Data Set

This is gene-expression microarray data from the breast cancer prognosis study conducted in the Netherlands and reported in van de Vijver et al. [2002]. The data set consists of 295 training points of dimensionality 70 and two class labels. The feature vectors are normalized gene-expression profiles from cells harvested from 295 beast tumor samples in a retrospective study, meaning that patients were tracked over the years and their outcomes recorded. Using this clinical information, the authors labeled the tumor samples into two classes: the "good prognosis" group (label 1) were disease-free for at least five years after first treatment, whereas the "bad prognosis" group developed distant metastasis within the first five years. Of the 295 patients, 216 belong to the "good-prognosis" class, whereas the remaining 79 belong to the "poor- prognosis" class.

A8.4 Stacking Fault Energy Data Set

This data set contains the experimentally recorded values of the stacking fault energy (SFE) in austenitic stainless steel specimens with different chemical compositions; see Yonezawa et al. [2013]. The SFE is a microscopic property related to the resistance of austenitic steels. High-SFE steels are less likely to fracture under strain and may be desirable in certain applications. The data set contains 17 features corresponding to the atomic element content of 473 steel specimens and the continuous-valued measured SFE for each.

A8.5 Soft Magnetic Alloy Data Set

This is a data set on Fe-based nanocrystalline soft magnetic alloys, which is part of on-going work [Wang et al., 2020]. This data set records the atomic composition and processing parameters along with several different electromagnetic properties for a large number of magnetic alloys. We will be particularly interested in the magnetic coercivity as the property to be predicted. Larger values of coercivity mean that the magnetized material has a wider hysteresis curve and can withstand larger magnetic external fields without losing its own magnetization. By contrast, small values of coercivity mean that a material can lose its magnetization quickly. Large-coercivity materials are therefore ideal to make permanent magnets, for example.

A8.6 Ultrahigh Carbon Steel Data Set

This is the Carnegie Mellon University Ultrahigh Carbon Steel (CMU-UHCS) dataset [Hecht et al., 2017; DeCost et al., 2017]. This data set consists of 961 high-resolution 645×484 images of steel samples subjected to a variety of heat treatments. The images are *micrographs* obtained by scanning electron microscopy (SEM) at several different magnifications. There are a total of seven different labels, corresponding to different phases of steel resulting from different thermal processing (number of images in parenthesis): spheroidite (374), carbide network (212), pearlite (124), pearlite + spheroidite (107), spheroidite+Widmanstätten (81), martensite (36), and pearlite+Widmanstätten (27). The main goal is to be able to predict the label of a new steel sample given its micrograph.

List of Symbols

$\mathbf{X} = (X_1, \ldots, X_d) \in R^d$	feature vector
$Y \in \{0, 1\}$	target
$c_i = P(Y = i)$, $i = 0, 1$	class prior probabilities
$P_{\mathbf{X}Y}$	feature-target distribution
$p(\mathbf{x})$	feature vector density (if it exists)
$p_i(\mathbf{x}) = p(\mathbf{x} \mid Y = i)$, $i = 0, 1$	class-conditional densities (if they exist)
$\eta(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$	posterior-probability function
$\psi : R^d \to \{0, 1\}$	classifier
$\varepsilon = \varepsilon[\psi] = P(\psi(\mathbf{X}) \neq Y)$	classifier error rate
ψ^*, ε^*	Bayes classifier and Bayes error
$\varepsilon^i = P(\psi(\mathbf{X}) = 1 - i \mid Y = i)$, $i = 0, 1$	population-specific true error rates
$D_n : R^d \to R$	sample-based discriminant
μ_i, $i = 0, 1$	class means
σ_i^2, $i = 0, 1$ class variances	
Σ_i, $i = 0, 1$	class covariance matrices
$\Phi(x) = (1/2\pi) \int_{-\infty}^{x} e^{-u^2} du$	cdf of a $N(0, 1)$ Gaussian random variable
$S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$	sample training data
n, n_0, $n_1 = n - n_0$	total and population-specific sample sizes
$\Psi_n : S_n \mapsto \psi_n$	classification rule
$\psi_n : R^d \to \{0, 1\}$	classifier designed from training data
$\varepsilon_n = \varepsilon[\psi_n] = P(\psi_n(\mathbf{X}) \neq Y \mid S_n)$	error rate of sample-based classifier

U. Braga-Neto, *Fundamentals of Pattern Recognition and Machine Learning*,
https://doi.org/10.1007/978-3-030-27656-0

$\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$	parameter vector and estimator
$k(\mathbf{x}, \mathbf{x}')$	kernel
λ_i, $i = 1, \ldots, d$	Lagrange multipliers
L_P, L_D	primal and dual Lagrangians
$\mathcal{C}$	set of classification rules
$V_{\mathcal{C}}$, $\mathcal{S}(\mathcal{C}, n)$	VC dimension and shatter coefficients
p_i, q_i, U_i, V_i, $i = 1, \ldots, b$	population-specific bin probabilities and counts
$\Xi_n : (\Psi_n, S_n, \xi) \mapsto \hat{\varepsilon}_n$	error estimation rule
$\hat{\varepsilon}_n$	error estimator for mixture sample
$\mathrm{Bias}(\hat{\varepsilon}_n)$, $\mathrm{Var}_{\mathrm{dev}}(\hat{\varepsilon}_n)$, $\mathrm{RMS}(\hat{\varepsilon}_n)$	bias, deviation variance, root mean square error
$S_m = \{(\mathbf{X}_i^t, Y_i^t); i = 1, \ldots, m\}$	independent test sample
$\hat{\varepsilon}_{n,m}$	test-set error estimator
$\hat{\varepsilon}_n^r$	resubstitution error estimator
$\hat{\varepsilon}_n^{\mathrm{cv}(k)}$	k-fold cross-validation error estimator
$\hat{\varepsilon}_n^l$	leave-one-out error estimator
$L[f]$	regression error of f
$\mathcal{F}$	σ-algebra
$\mathcal{B}$	Borel σ-algebra
μ,ν	measures
λ	Lebesgue measure
$X_n \xrightarrow{a.s.} X$	almost-sure convergence (with probability 1) of X_n to X
$X_n \xrightarrow{L^p} X$	L^p convergence of X_n to X
$X_n \xrightarrow{P} X$	convergence of X_n to X in probability
$X_n \xrightarrow{D} X$	convergence of X_n to X in distribution

Bibliography

Afsari, B., Braga-Neto, U., and Geman, D. (2014). Rank discriminants for predicting phenotypes from rna expression. *Annals of Applied Statistics*, 8(3):1469–1491.

Aitchison, J. and Dunsmore, I. (1975). *Statistical Predication Analysis*. Cambridge University Press, Cambridge, UK.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (2002). *Molecular Biology of the Cell*. Garland, 4th edition.

Alvarez, S., Diaz-Uriarte, R., Osorio, A., Barroso, A., Melchor, L., Paz, M., Honrado, E., Rodriguez, R., Urioste, M., Valle, L., Diez, O., Cigudosa, J., Dopazo, J., Esteller, M., and Benitez, J. (2005). A predictor based on the somatic genomic changes of the brca1/brca2 breast cancer tumors identifies the non-brca1/brca2 tumors with brca1 promoter hypermethylation. *Clin Cancer Res*, 11(3):1146–1153.

Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl. Acad. Sci.*, 99(10):6562–6566.

Anderson, T. (1951). Classification by multivariate analysis. *Psychometrika*, 16:31–50.

Anderson, T. (1973). An asymptotic expansion of the distribution of the studentized classification statistic W. *The Annals of Statistics*, 1:964–972.

Anderson, W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2nd edition.

Bartlett, P., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48:85–113.

Bertsekas, D. (1995). *Nonlinear Programming*. Athena Scientific.

Billingsley, P. (1995). *Probability and Measure*. John Wiley, New York City, New York, third edition.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

U. Braga-Neto, *Fundamentals of Pattern Recognition and Machine Learning*,
https://doi.org/10.1007/978-3-030-27656-0

Boser, B., Guyon, M., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Workshop on Computational Learning Theory*.

Bowker, A. (1961). A representation of hotelling's t^2 and anderson's classification statistic w in terms of simple statistics. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, pages 285–292. Stanford University Press.

Bowker, A. and Sitgreaves, R. (1961). An asymptotic expansion for the distribution function of the w-classification statistic. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, pages 292–310. Stanford University Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Braga-Neto, U. (2007). Fads and fallacies in the name of small-sample microarray classification. *IEEE Signal Processing Magazine*, 24(1):91–99.

Braga-Neto, U., Arslan, E., Banerjee, U., and Bahadorinejad, A. (2018). Bayesian classification of genomic big data. In Sedjic, E. and Falk, T., editors, *Signal Processing and Machine Learning for Biomedical Big Data*. Chapman and Hall/CRC Press.

Braga-Neto, U. and Dougherty, E. (2004). Bolstered error estimation. *Pattern Recognition*, 37(6):1267–1281.

Braga-Neto, U. and Dougherty, E. (2015). *Error Estimation for Pattern Recognition*. Wiley, New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Bryson, A. and Ho, Y.-C. (1969). *Applied Optimal Control: Optimization, Estimation, and Control*. Blaisdell Publishing Company.

Buduma, N. and Locascio, N. (2017). *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. O'Reilly Media, Inc.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition.

Castro, R. (2020). Statistical Learning Theory Lecture Notes. Accessed: Jun 12, 2020. https://www.win.tue.nl/∼rmcastro/2DI70/files/2DI70_Lecture_Notes.pdf.

Chapelle, O., Scholkopf, B., and Zien, A., editors (2010). *Semi-Supervised Learning*. MIT Press.

Cherkassky, V. and Ma, Y. (2003). Comparison of model selection for regression. *Neural computation*, 15(7):1691–1714.

Cherkassky, V., Shao, X., Mulier, F. M., and Vapnik, V. N. (1999). Model complexity control for regression using vc generalization bounds. *IEEE transactions on Neural Networks*, 10(5):1075–1089.

Chernick, M. (1999). *Bootstrap Methods: A Practitioner's Guide*. John Wiley & Sons, New York.

Chung, K. L. (1974). *A Course in Probability Theory, Second Edition*. Academic Press, New York City, New York.

Cover, T. (1969). Learning in pattern recognition. In Watanabe, S., editor, *Methodologies of Pattern Recognition*, pages 111–132. Academic Press, New York, NY.

Cover, T. and Hart, P. (1967). Nearest-neighbor pattern classification. *IEEE Trans. on Information Theory*, 13:21–27.

Cover, T. and van Campenhout, J. (1977). On the possible orderings in the measurement selection problem. *IEEE Trans. on Systems, Man, and Cybernetics*, 7:657–661.

Cover, T. M. (1974). The best two independent measurements are not the two best. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-4(1):116–117.

Cramér, H. (1999). *Mathematical methods of statistics*, volume 43. Princeton university press.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley, New York City, New York.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dalton, L. and Dougherty, E. (2011a). Application of the bayesian mmse error estimator for classification error to gene-expression microarray data. *IEEE Transactions on Signal Processing*, 27(13):1822–1831.

Dalton, L. and Dougherty, E. (2011b). Bayesian minimum mean-square error estimation for classification error part I: Definition and the bayesian mmse error estimator for discrete classification. *IEEE Transactions on Signal Processing*, 59(1):115–129.

Dalton, L. and Dougherty, E. (2011c). Bayesian minimum mean-square error estimation for classification error part II: Linear classification of gaussian models. *IEEE Transactions on Signal Processing*, 59(1):130–144.

Dalton, L. and Dougherty, E. (2012a). Exact mse performance of the bayesian mmse estimator for classification error part i: Representation. *IEEE Transactions on Signal Processing*, 60(5):2575–2587.

Dalton, L. and Dougherty, E. (2012b). Exact mse performance of the bayesian mmse estimator for classification error part ii: Performance analysis and applications. *IEEE Transactions on Signal Processing*, 60(5):2588–2603.

Dalton, L. and Dougherty, E. (2013). Optimal classifiers with minimum expected error within a bayesian framework – part I: Discrete and gaussian models. *Pattern Recognition*, 46(5):1301–1314.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

De Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software*, 31(3).

DeCost, B. L., Francis, T., and Holm, E. A. (2017). Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures. *Acta Materialia*, 133:30–40.

Dempster, A. D., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Devroye, L. and Wagner, T. (1976). Nonparametric discrimination and density estimation. Technical Report 183, Electronics Research Center, University of Texas, Austin, TX.

Dougherty, E. R. and Brun, M. (2004). A probabilistic theory of clustering. *Pattern Recognition*, 37(5):917–925.

Duda, R., Hart, P., and Stork, G. (2001). *Pattern Classification*. John Wiley & Sons, New York, 2nd edition.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7):research0036–1.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *Annals of Statistics*, 7:1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Elashoff, J. D., Elashoff, R., and COLDMAN, G. (1967). On the choice of variables in classification problems with dichotomous variables. *Biometrika*, 54(3-4):668–670.

Esfahani, S. and Dougherty, E. (2014). Effect of separate sampling on classification accuracy. *Bioinformatics*, 30(2):242–250.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*. Wiley, New York, 3rd edition.

Fei-Fei, L., Deng, J., Russakovski, O., Berg, A., and Li, K. (2010). ImageNet Summary and Statistics. http://www.image-net.org/about-stats. Accessed: Jan 2, 2020.

Fisher, R. (1935). The fiducial argument in statistical inference. *Ann. Eugen.*, 6:391–398.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7(2):179–188.

Fix, E. and Hodges, J. (1951). Nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX. Project Number 21-49-004.

Foley, D. (1972). Considerations of sample and feature size. *IEEE Transactions on Information Theory*, IT-18(5):618–626.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *Proc. Third Annual Workshop on Computational Learning Theory*, pages 202–216.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *Journal of the Royal Statistical Society: Series B*, 26(1):69–76.

Geman, D., d'Avignon, C., Naiman, D., Winslow, R., and Zeboulon, A. (2004). Gene expression comparisons for class prediction in cancer studies. In *Proceedings of the 36th Symposium on the Interface: Computing Science and Statistics*, Baltimore, MD.

Girosi, F. and Poggio, T. (1989). Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4):465–469.

Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, 29(2):241–256.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, 10:211–222.

Groenen, P. J., van de Velden, M., et al. (2016). Multidimensional scaling by majorization: A review. *Journal of Statistical Software*, 73(8):1–26.

Hamamoto, Y., Uchimura, S., Matsunra, Y., Kanaoka, T., and Tomita, S. (1990). Evaluation of the branch and bound algorithm for feature selection. *Pattern Recognition Letters*, 11:453–456.

Hanczar, B., Hua, J., and Dougherty, E. (2007). Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007. Article ID 38473, 12 pages.

Hand, D. (1986). Recent advances in error rate estimation. *Pattern Recognition Letters*, 4:335–346.

Harter, H. (1951). On the distribution of wald's classification statistics. *Ann. Math. Statist.*, 22:58–67.

Hassan, M. (2018). VGG16 convolutional network for classification and detection. https://neurohive.io/en/popular-networks/vgg16/. Accessed: Jan 1, 2020.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hecht, M. D., Picard, Y. N., and Webler, B. A. (2017). Coarsening of inter-and intra-granular proeutectoid cementite in an initially pearlitic 2c-4cr ultrahigh carbon steel. *Metallurgical and Materials Transactions A*, 48(5):2320–2335.

Hills, M. (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):1–31.

Hirst, D. (1996). Error-rate estimation in multiple-group linear discriminant analysis. *Technometrics*, 38(4):389–399.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press, New York, NY.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hua, J., Tembe, W., and Dougherty, E. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42:409–424.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, IT-14(1):55–63.

Izmirlian, G. (2004). Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann. NY. Acad. Sci.*, 1020:154–174.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):153–158.

Jain, A. K., Dubes, R. C., et al. (1988). *Algorithms for clustering data*, volume 6. Prentice hall Englewood Cliffs, NJ.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jazwinski, A. H. (2007). *Stochastic processes and filtering theory*. Courier Corporation.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, UK, 3rd edition.

Jiang, X. and Braga-Neto, U. (2014). A naive-bayes approach to bolstered error estimation in high-dimensional spaces. Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2014), Atlanta, GA.

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier.

John, S. (1961). Errors in discrimination. *Ann. Math. Statist.*, 32:1125–1144.

Kaariainen, M. (2005). Generalization error bounds using unlabeled data. In *Proceedings of COLT'05*.

Kaariainen, M. and Langford, J. (2005). A comparison of tight generalization bounds. In *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany.

Kabe, D. (1963). Some results on the distribution of two random matrices used in classification procedures. *Ann. Math. Statist.*, 34:181–185.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kim, S., Dougherty, E., Barrera, J., Chen, Y., Bittner, M., and Trent, J. (2002). Strong feature sets from small samples. *Computational Biology*, 9:127–146.

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS microbiology reviews*, 35(2):343–359.

Kohane, I., Kho, A., and Butte, A. (2003). *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1143, Montreal, CA.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kudo, M. and Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41.

Kullback, S. (1968). *Information Theory and Statistics*. Dover, New York.

Lachenbruch, P. (1965). *Estimation of error rates in discriminant analysis*. PhD thesis, University of California at Los Angeles, Los Angeles, CA.

Lachenbruch, P. and Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Linnaeus, C. (1758). *Systema naturae*. Impensis Laurentii Salvii, 10th edition.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680.

Loève, M. (1977). *Probability Theory I*. Springer.

Lorentz, G. G. (1976). The 13th problem of hilbert. In *Proceedings of Symposia in Pure Mathematics*, volume 28, pages 419–430. American Mathematical Society.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239.

Lugosi, G. and Pawlak, M. (1994). On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2):475–481.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Marguerat, S. and Bahler, J. (2010). Rna-seq: from technology to biology. *Cellular and molecular life science*, 67(4):569–579.

Martins, D., Braga-Neto, U., Hashimoto, R., Bittner, M., and Dougherty, E. (2008). Intrinsically multivariate predictive genes. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):424–439.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McFarland, H. and Richards, D. (2001). Exact misclassification probabilities for plug-in normal quadratic discriminant functions. i. the equal-means case. *Journal of Multivariate Analysis*, 77:21–53.

McFarland, H. and Richards, D. (2002). Exact misclassification probabilities for plug-in normal quadratic discriminant functions. ii. the heterogeneous case. *Journal of Multivariate Analysis*, 82:299–330.

McLachlan, G. (1976). The bias of the apparent error in discriminant analysis. *Biometrika*, 63(2):239–244.

McLachlan, G. (1987). Error rate estimation in discriminant analysis: recent advances. In Gupta, A., editor, *Advances in Multivariate Analysis*. D. Reidel, Dordrecht.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley-Interscience, New York.

Minka, T. (1998). Expectation maximization as lower bound maximization. Technical report, Microsoft Research. Tutorial published on the web at http://www-white.media.mit.edu/tpminka/papers/em.html.

Moran, M. (1975). On the expectation of errors of allocation associated with a linear discriminant function. *Biometrika*, 62(1):141–148.

Murphy, K. (2012a). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murphy, K. P. (2012b). *Machine learning: a probabilistic perspective*. MIT press.

Narendra, P. and Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computers*, 26(9):917–922.

Nascimento, E., Abath, F., Calzavara, C., Gomes, A., Acioli, B., Brito, C., Cordeiro, M., Silva, A., Andrade, C. M. R., Gil, L., and Junior, U. B.-N. E. M. (2009). Gene expression profiling during early acute febrile stage of dengue infection can predict the disease outcome. *PLoS ONE*, 4(11):e7892. doi:10.1371/journal.pone.0007892.

Nilsson, R., Peña, J. M., Björkegren, J., and Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8(Mar):589–612.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Nualart, D. (2004). Kolmogorov and probability theory. *Arbor*, 178(704):607–619.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, 34:1286–1301. Correction: Ann. Math. Statist., 39:1358–1359, 1968.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

Poor, V. and Looze, D. (1981). Minimax state estimation for linear stochastic systems with noise uncertainty. *IEEE Transactions on Automatic Control*, AC-26(4):902–906.

Rajan, K., editor (2013). *Informatics for Materials Science and Engineering*. Butterworth-Heinemann, Waltham, MA.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Raudys, S. (1972). On the amount of a priori information in designing the classification algorithm. *Technical Cybernetics*, 4:168–174. in Russian.

Raudys, S. (1978). Comparison of the estimates of the probability of misclassification. In *Proc. 4th Int. Conf. Pattern Recognition*, pages 280–282, Kyoto, Japan.

Raudys, S. and Jain, A. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):4–37.

Raudys, S. and Young, D. (2004). Results in statistical discriminant analysis: a review of the former soviet union literature. *Journal of Multivariate Analysis*, 89:1–35.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2nd edition.

Rogers, W. and Wagner, T. (1978). A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics*, 6:506–514.

Rosenblatt, F. (1957). The perceptron – a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, Buffalo, NY.

Rosenthal, J. (2006). *A First Look At Rigorous Probability Theory*. World Scientific Publishing, Singapore, 2nd edition.

Ross, S. (1994). *A first course in probability*. Macmillan, New York, 4th edition.

Ross, S. (1995). *Stochastic Processes*. Wiley, New York, 2nd edition.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sayre, J. (1980). The distributions of the actual error rates in linear discriminant analysis. *Journal of the American Statistical Association*, 75(369):201–205.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns via a complementary DNA microarray. *Science*, 270:467–470.

Schiavo, R. and Hand, D. (2000). Ten more years of error rate research. *International Statistical Review*, 68(3):295–310.

Schroeder, M. (2009). *Fractals, chaos, power laws: Minutes from an infinite paradise*. Dover.

Sima, C., Attoor, S., Braga-Neto, U., Lowey, J., Suh, E., and Dougherty, E. (2005a). Impact of error estimation on feature-selection algorithms. *Pattern Recognition*, 38(12):2472–2482.

Sima, C., Braga-Neto, U., and Dougherty, E. (2005b). Bolstered error estimation provides superior feature-set ranking for small samples. *Bioinformatics*, 21(7):1046–1054.

Sima, C. and Dougherty, E. (2006). Optimal convex error estimators for classification. *Pattern Recognition*, 39(6):1763–1780.

Sima, C., Vu, T., Braga-Neto, U., and Dougherty, E. (2014). High-dimensional bolstered error estimation. *Bioinformatics*, 27(21):3056–3064.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sitgreaves, R. (1951). On the distribution of two random matrices used in classification procedures. *Ann. Math. Statist.*, 23:263–270.

Sitgreaves, R. (1961). Some results on the distribution of the W-classification. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, pages 241–251. Stanford University Press.

Smith, C. (1947). Some examples of discrimination. *Annals of Eugenics*, 18:272–282.

Snapinn, S. and Knoke, J. (1985). An evaluation of smoothed classification error-rate estimators. *Technometrics*, 27(2):199–206.

Snapinn, S. and Knoke, J. (1989). Estimation of error rates in discriminant analysis with selection of variables. *Biometrics*, 45:289–299.

Stark, H. and Woods, J. W. (1986). *Probability, random processes, and estimation theory for engineers*. Prentice-Hall, Inc.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9:465–474.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5:595–645.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:111–147.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT press Cambridge.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6):2907–2912.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904.

Tanaseichuk, O., Borneman, J., and Jiang, T. (2013). Phylogeny-based classification of microbial communities. *Bioinformatics*, 30(4):449–456.

Teichroew, D. and Sitgreaves, R. (1961). Computation of an empirical sampling distribution for the w-classification statistic. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, pages 285–292. Stanford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99:6567–6572.

Toussaint, G. (1971). Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory*, 17(5):618.

Toussaint, G. (1974). Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, IT-20(4):472–479.

Toussaint, G. and Donaldson, R. (1970). Algorithms for recognizing contour-traced hand-printed characters. *IEEE Transactions on Computers*, 19:541–546.

Toussaint, G. and Sharpe, P. (1974). An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis. *IEEE Transactions on Information Theory*, IT-20(4):472–479.

Tutz, G. (1985). Smoothed additive estimators for non-error rates in multiple discriminant analysis. *Pattern Recognition*, 18(2):151–159.

van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Astma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Vitushkin, A. (1954). On hilberts thirteenth problem. *Dokl. Akad. Nauk SSSR*, 95(4):701–704.

Vu, T., Braga-Neto, U., and Dougherty, E. (2008). Preliminary study on bolstered error estimation in high-dimensional spaces. In *Proceedings of GENSIPS'2008 - IEEE International Workshop on Genomic Signal Processing and Statistics*. Phoenix, AZ.

Vu, T., Sima, C., Braga-Neto, U., and Dougherty, E. (2014). Unbiased bootstrap error estimation for linear discrimination analysis. *EURASIP Journal on Bioinformatics and Systems Biology*, 2014:15.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.*, 15:145–162.

Wang, Y., Tian, Y., Kirk, T., Laris, O., Ross Jr, J. H., Noebe, R. D., Keylin, V., and Arróyave, R. (2020). Accelerated design of fe-based soft magnetic materials using machine learning and stochastic optimization. *Acta Materialia*, 194:144–155.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Webb, A. (2002). *Statistical Pattern Recognition*. John Wiley & Sons, New York, 2nd edition.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behaviorial Sciences*. PhD thesis, Harvard University, Cambridge, MA.

Wolpert, D. (2001). The supervised learning no-free-lunch theorems. In *World Conference on Soft Computing*.

Wyman, F., Young, D., and Turner, D. (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition*, 23(7):775–783.

Xiao, Y., Hua, J., and Dougherthy, E. (2007). Quantification of the impact of feature selection on cross-validation error estimation precision. *EURASIP J. Bioinformatics and Systems Biology*.

Xie, S. and Braga-Neto, U. M. (2019). On the bias of precision estimation under separate sampling. *Cancer informatics*, 18:1–9.

Xu, Q., Hua, J., Braga-Neto, U., Xiong, Z., Suh, E., and Dougherty, E. (2006). Confidence intervals for the true classification error conditioned on the estimated error. *Technology in Cancer Research and Treatment*, 5(6):579–590.

Yonezawa, T., Suzuki, K., Ooki, S., and Hashimoto, A. (2013). The effect of chemical composition and heat treatment conditions on stacking fault energy for fe-cr-ni austenitic stainless steel. *Metallurgical and Materials Transactions A*, 44A:5884–5896.

Zhou, X. and Mao, K. (2006). The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms. *Bioinformatics*, 22:2507–2515.

Zollanvari, A., Braga-Neto, U., and Dougherty, E. (2009a). On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. *Pattern Recognition*, 42(11):2705–2723.

Zollanvari, A., Braga-Neto, U., and Dougherty, E. (2010). Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis. *IEEE Transactions on Information Theory*, 56(2):784–804.

Zollanvari, A., Braga-Neto, U., and Dougherty, E. (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Transactions on Signal Processing*, 59(9):1–18.

Zollanvari, A., Braga-Neto, U., and Dougherty, E. (2012). Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate heteroskedastic gaussian model. *Pattern Recognition*, 45(2):908–917.

Zollanvari, A., Cunningham, M. J., Braga-Neto, U., and Dougherty, E. R. (2009b). Analysis and modeling of time-course gene-expression profiles from nanomaterial-exposed primary human epidermal keratinocytes. *BMC Bioinformatics*, 10(11):S10.

Zollanvari, A. and Dougherty, E. (2014). Moments and root-mean-square error of the bayesian mmse estimator of classification error in the gaussian model. *Pattern Recognition*, 47(6):2178–2192.

Zolman, J. (1993). *Biostatistics: Experimental Design and Statistical Inference*. Oxford University Press, New York, NY.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Index

U. Braga-Neto, *Fundamentals of Pattern Recognition and Machine Learning*,
https://doi.org/10.1007/978-3-030-27656-0