1

Nonparametric Statistical Models

In this chapter we introduce and motivate the statistical models that will be considered in this book. Some of the materials depend on basic facts developed in subsequent chapters – mostly the basic Gaussian process and Hilbert space theory. This will be hinted at when necessary.

Very generally speaking, a *statistical model* for a random observation Y is a family

$$\{P_f : f \in \mathcal{F}\}$$

of probability distributions P_f, each of which is a candidate for having generated the observation Y. The parameter f belongs to the *parameter space* $\mathcal{F}$. The problem of *statistical inference* on f, broadly speaking, can be divided into three intimately connected problems of using the observation Y to

(a) *Estimate* the parameter f by an estimator $T(Y)$,
(b) *Test hypotheses* on f based on test functions $\Psi(Y)$ and/or
(c) *Construct confidence sets* $C(Y)$ that contain f with high probability.

To interpret inferential results of these kinds, we will typically need to specify a distance, or loss function on $\mathcal{F}$, and for a given model, different loss functions may or may not lead to very different conclusions.

The statistical models we will introduce in this chapter are, on the one hand, conceptually closely related to each other in that the parameter space $\mathcal{F}$ is infinite or high dimensional and the loss functions relevant to the analysis of the performance of statistical procedures are similar. On the other hand, these models are naturally divided by the different probabilistic frameworks in which they occur – which will be either a *Gaussian noise model* or an *independent sampling model*. These frameworks are asymptotically related in a fundamental way (see the discussion after Theorem 1.2.1). However, the most effective probabilistic techniques available are based on a direct, nonasymptotic analysis of the Gaussian or product probability measures that arise in the relevant sampling context and hence require a separate treatment.

Thus, while many of the statistical intuitions are common to both the sampling and the Gaussian noise models and in fact inform each other, the probabilistic foundations of these models will be laid out independently.

1.1 Statistical Sampling Models

Let X be a random experiment with associated sample space $\mathcal{X}$. We take the mathematical point of view of probability theory and model X as a random variable, that is, as a measurable mapping defined on some underlying probability space that takes values in the measurable space $(\mathcal{X},\mathcal{A})$, where $\mathcal{A}$ is a σ-field of subsets of $\mathcal{X}$. The law of X is described by the probability measure P on $\mathcal{A}$. We may typically think of $\mathcal{X}$ equal to $\mathbb{R}^d$ or a measurable subset thereof, equipped with its Borel σ-field $\mathcal{A}$.

The perhaps most basic problem of statistics is the following: consider repeated outcomes of the experiment X, that is, a random sample of independent and identically distributed (i.i.d.) copies $X_1,\dots,X_n$ from X. The joint distribution of the X_i equals the product probability measure $P^n = \otimes_{i=1}^n P$ on $(\mathcal{X}^n,\mathcal{A}^n)$. The goal is to recover P from the n observations. 'Recovering P' can mean many things. Classical statistics has been concerned mostly with models where P is explicitly parameterised by a finite-dimensional parameter, such as the mean and variance of the normal distribution, or the 'parameters' of the usual families of statistical distributions (gamma, beta, exponential, Poisson, etc.). Recovering P then simply means to use the observations to make inferences on the unknown parameter, and the fact that this parameter is finite dimensional is crucial for this traditional paradigm of statistical inference, in particular, for the famous likelihood principle of R. A. Fisher. In this book, we will follow the often more realistic assumption that no such parametric assumptions are made on P. For most sample spaces $\mathcal{X}$ of interest, this will naturally lead to models that are infinite dimensional, and we will investigate how the theory of statistical inference needs to be developed in this situation.

1.1.1 Nonparametric Models for Probability Measures

In its most elementary form, without imposing any parameterisations on P, we can simply consider the problem of making inferences on the unknown probability measure P based on the sample. Natural loss functions arise from the usual metrics on the space of probability measures on $\mathcal{X}$, such as the total variation metric

$$\|P-Q\|_{TV} = \sup_{A\in\mathcal{A}} |P(A)-Q(A)|$$

or weaker metrics that generate the topology of weak convergence of probability measures on $\mathcal{X}$. For instance, if $\mathcal{X}$ itself is endowed with a metric d, we could take the bounded Lipschitz metric

$$\beta_{(\mathcal{X},d)}(P,Q) = \sup_{f\in BL(1)} \left| \int_{\mathcal{X}} f\,(dP-dQ) \right|$$

for weak convergence of probability measures, where

$$BL(M) = \left\{ f:\mathcal{X}\to\mathbb{R},\ \sup_{x\in\mathcal{X}} |f(x)| + \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)} \le M \right\}, \quad 0<M<\infty.$$

If $\mathcal{X}$ has some geometric structure, we can consider more intuitive loss functions. For example, if $\mathcal{X}=\mathbb{R}$, we can consider the cumulative distribution function

$$F(x) = P(X\le x), \quad x\in\mathbb{R},$$

or, if X takes values in $\mathbb{R}^d$, its multivariate analogue. A natural distance function on distribution functions is simply the supremum-norm metric ('Kolmogorov distance')

$$\|F_P - F_Q\|_\infty = \sup_{x\in\mathbb{R}} |F_P(x) - F_Q(x)|.$$

Since the indicators $\{1_{(-\infty,x]} : x \in \mathbb{R}\}$ generate the Borel σ-field of $\mathbb{R}$, we see that, on $\mathbb{R}$, the statistical parameter P is characterised entirely by the functional parameter F, and vice versa. The parameter space is thus the infinite-dimensional space of all cumulative distribution functions on $\mathbb{R}$.

Often we will know that P has some more structure, such as that P possesses a probability-density function $f : \mathbb{R} \to [0,\infty)$, which itself may have further properties that will be seen to influence the complexity of the statistical problem at hand. For probability-density functions, a natural loss function is the L^1-distance

$$\|f_P - f_Q\|_1 = \int_{\mathbb{R}} |f_P(x) - f_Q(x)|dx$$

and in some situations also other L^p-type and related loss functions. Although in some sense a subset of the other, the class of probability densities is more complex than the class of probability-distribution functions, as it is not described by monotonicity constraints and does not consist of functions bounded in absolute value by 1. In a heuristic way, we can anticipate that estimating a probability density is harder than estimating the distribution function, just as the preceding total variation metric is stronger than any metric for weak convergence of probability measures (on nontrivial sample spaces $\mathcal{X}$). In all these situations, we will see that the theory of statistical inference on the parameter f significantly departs from the usual finite-dimensional setting.

Instead of P, a particular functional $\Phi(P)$ may be the parameter of statistical interest, such as the moments of P or the quantile function F^{-1} of the distribution function F – examples for this situation are abundant. The nonparametric theory is naturally compatible with such functional estimation problems because it provides the direct plug-in estimate $\Phi(T)$ based on an estimator T for P. Proving closeness of T to P in some strong loss function then gives access to 'many' continuous functionals Φ for which $\Phi(T)$ will be close to $\Phi(P)$, as we shall see later in this book.

1.1.2 Indirect Observations

A common problem in statistical sampling models is that some systematic measurement errors are present. A classical problem of this kind is the statistical regression problem, which will be introduced in the next section. Another problem, which is more closely related to the sampling model from earlier, is where one considers observations in $\mathbb{R}^d$ of the form

$$Y_i = X_i + \varepsilon_i, \quad i = 1,\dots,n, \tag{1.1}$$

where the X_i are i.i.d. with common law P_X, and the ε_i are random 'error' variables that are independent of the X_i and have law P_ε. The law P_ε is assumed to be known to the observer – the nature of this assumption is best understood by considering examples: the attempt is to model situations in which a scientist, for reasons of cost, complexity or lack of precision of the involved measurement device, is forced to observe Y_i instead of the

realisations X_i of interest. The observer may, however, have very concrete knowledge of the source of the error, which could, for example, consist of light emissions of the Milky Way interfering with cosmic rays from deeper space, an erratic optical device through which images are observed (e.g., a space telescope which cannot be repaired except at very high cost) or transmissions of signals through a very busy communication channel. Such situations of implicit measurements are encountered frequently in the applied sciences and are often called *inverse problems*, as one wishes to 'undo' the errors inflicted on the signal in which one is interested. The model (1.1) gives a simple way to model the main aspects of such statistical inverse problems. It is also known as the *deconvolution model* because the law of the Y_i equals

$$P_Y = P_X * P_\varepsilon,$$

the convolution of the two probability measures P_X, P_ε, and one wishes to 'deconvolve' P_ε.

As earlier, we will be interested in inference on the underlying distribution P_X of the signal X when the statistical model for P_X is infinite dimensional. The loss functions in this problem are thus typically the same as in the preceding subsection.

1.2 Gaussian Models

The randomness in the preceding sampling model was encoded in a general product measure P^n describing the joint law of the observations. Another paradigm of statistical modelling deals with situations in which the randomness in the model is described by a Gaussian (normal) distribution. This paradigm naturally encompasses a variety of nonparametric models, where the infinite-dimensional character of the problem does not necessarily derive from the probabilistic angle but from a functional relationship that one wishes to model.

1.2.1 Basic Ideas of Regression

Perhaps the most natural occurrence of a statistical model in the sciences is the one in which observations, modelled here as numerical values or vectors, say, (Y_i, x_i), arise according to a functional relationship

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1.2}$$

where n is the number of observations (sample size), f is some function of the x_i and the ε_i are random noise. By 'random noise', we may mean here either a probabilistic model for certain measurement errors that we believe to be intrinsic to our method of making observations, or some innate stochastic nature of the way the Y_i are generated from the $f(x_i)$. In either case, we will model the ε_i as random variables in the sense of axiomatic probability theory – the question of the genuine physical origin of this random noise will not concern us here. It is sometimes natural to assume also that the x_i are realisations of random variables X_i – we can either take this into account explicitly in our analysis or make statements conditional on the observed values $X_i = x_i$.

The function f often will be unknown to the observer of observations (Y_i, x_i), and the goal is to recover f from the (Y_i, x_i). This may be of interest for various reasons, for instance, for predicting new values Y_{n+1} from $f(x_{n+1})$ or to gain quantitative and qualitative understanding of the functional relationship $Y_i = f(x_i)$ under consideration.

In the preceding context, a statistical model in the broad sense is an a priori specification of both a parameter space for the functions f that possibly could have generated (1.2) and a family of probability measures that describes the possible distributions of the random variables ε_i. By 'a priori', we mean here that this is done independently of (e.g., before) the observational process, reflecting the situation of an experimentalist.

A systematic use and study of such models was undertaken in the early nineteenth century by Carl Friedrich Gauss, who was mostly interested in predicting astronomical observations. When the model is translated into the preceding formalisation, Gauss effectively assumed that the x_i are vectors $(x_{i1},\dots,x_{ip})^T$ and thought of f as a linear function in that vector, more precisely,

$$f(x_i) = x_{i1}\theta_i + \dots x_{ip}\theta_p, \quad i = 1,\dots,n,$$

for some real-valued parameters $\theta_j, j = 1,\dots,p$. The parameter space for f is thus the Euclidean space $\mathbb{R}^p$ expressed through all such linear mappings. In Gauss's time, the assumption of linearity was almost a computational necessity.

Moreover, Gauss modelled the random noise ε_i as independent and identically distributed samples from a normal distribution $N(0,\sigma^2)$ with some variance σ^2. His motivation behind this assumption was twofold. First, it is reasonable to assume that $E(\varepsilon_i) = 0$ for every i. If this expectation were nonzero, then there would be some deterministic, or 'systematic', measurement error $e_i = E(\varepsilon_i)$ of the measurement device, and this could always be accommodated in the functional model by adding a constant $x_{10} = \dots = x_{n0} = 1$ to the preceding linear relationship. The second assumption that ε_i has a normal distribution is deeper. If we think of each measurement error ε_i as the sum of many 'very small', or infinitesimal, independent measurement errors $\varepsilon_{ik}, k = 1,2,\dots$, then, by the central limit theorem, $\varepsilon_i = \sum_k \varepsilon_{ik}$ should be approximately normally distributed, regardless of the actual distribution of the ε_{ik}. By the same reasoning, it is typically natural to assume that the ε_i are also independent among themselves. This leads to what is now called the *standard Gaussian linear model*

$$Y_i = f(x_i) + \varepsilon_i \equiv \sum_{j=1}^{p} x_{ij}\theta_j + \varepsilon_i, \qquad \varepsilon_i \sim^{i.i.d.} N(0,\sigma^2), \quad i = 1,\dots,n, \tag{1.3}$$

which bears this name both because Gauss studied it and, since the $N(0,\sigma^2)$ distribution is often called the *Gaussian distribution*, because Gauss first made systematic use of it. The unknown parameter (θ,σ^2) varies in the $(p+1)$-dimensional parameter space

$$\Theta \times \Sigma = \mathbb{R}^p \times (0,\infty).$$

This model constitutes perhaps *the* classical example of a *finite-dimensional model*, which has been studied extensively and for which a fairly complete theory is available. For instance, when p is smaller than n, the least-squares estimator of Gauss finds the value $\hat\theta \in \mathbb{R}^p$ which solves the optimisation problem

$$\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2$$

and hence minimises the Euclidean distance of the vector $Y = (Y_1,\dots,Y_n)^T$ to the p-dimensional subspace spanned by the p vectors $(x_{1j},\dots,x_{nj})^T, j = 1,\dots,p$.

1.2.2 Some Nonparametric Gaussian Models

We now give a variety of models that generalise Gauss's ideas to infinite-dimensional situations. In particular, we will introduce the Gaussian white noise model, which serves as a generic surrogate for a large class of nonparametric models, including even non-Gaussian ones, through the theory of equivalence of experiments (discussed in the next section).

Nonparametric Gaussian Regression

Gauss's model and its theory basically consist of two crucial assumptions: one is that the ε_i are normally distributed, and the other is that the function f is linear. The former assumption was argued to be in some sense natural, at least in a measurement-error model (see also the remarks after Theorem 1.2.1 for further justification). The latter assumption is in principle quite arbitrary, particularly in times when computational power does not constrain us as much any longer as it did in Gauss's time. A nonparametric approach therefore attempts to assume as little structure of f as possible. For instance, by the *nonparametric regression model with fixed, equally spaced design on* $[0,1]$, we shall understand here the model

$$Y_i = f(x_i) + \varepsilon_i, \qquad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N(0,\sigma^2), \quad i = 1,\dots,n. \tag{1.4}$$

where f is any function defined on $[0,1]$. We are thus sampling the unknown function f at an equally spaced grid of $[0,1]$ that, as $n \to \infty$, grows dense in the interval $[0,1]$ as $n \to \infty$.

The model immediately generalises to bounded intervals $[a,b]$, to 'approximately' equally spaced designs $\{x_i : i = 1,\dots,n\} \subset [a,b]$ and to multivariate situations, where the x_i are equally spaced points in some hypercube. We note that the assumption that the x_i are equally spaced is important for the theory that will follow – this is natural as we cannot hope to make inference on f in regions that contain no or too few observations x_i.

Other generalisations include the *random design regression model*, in which the x_i are viewed as i.i.d. copies of a random variable X. One can then either proceed to argue conditionally on the realisations $X_i = x_i$, or one takes this randomness explicitly into account by making probability statements under the law of X and ε simultaneously. For reasonable design distributions, this will lead to results that are comparable to the fixed-design model – one way of seeing this is through the equivalence theory for statistical experiments (see after Theorem 1.2.1).

A priori it may not be reasonable to assume that f has any specific properties other than that it is a continuous or perhaps a differentiable function of its argument. Even if we would assume that f has infinitely many continuous derivatives the set of all such f would be infinite dimensional and could never be fully captured by a p-dimensional parameter space. We thus have to expect that the theory of statistical inference in this nonparametric model will be different from the one in Gauss's classical linear model.

The Gaussian White Noise Model

For the mathematical development in this book we shall work with a mathematical idealisation of the regression model (1.4) in continuous time, known as the *Gaussian white noise model*, and with its infinite sequence space analogue. While perhaps at first appearing more complicated than the discrete model, once constructed, it allows for a clean

and intuitive mathematical exposition that mirrors all the main ideas and challenges of the discrete case with no severe loss of generality.

Consider the following stochastic differential equation:

$$dY(t) \equiv dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t), \qquad t \in [0,1], \quad n \in \mathbb{N}, \tag{1.5}$$

where $f \in L^2 \equiv L^2([0,1])$ is a square integrable function on $[0,1]$, $\sigma > 0$ is a dispersion parameter and dW is a *standard Gaussian white noise process*. When we observe a realisation of (1.5), we shall say that we observe the function or signal f in Gaussian white noise, at the noise level, or a signal-to-noise ratio $\sigma/\sqrt{n}$. We typically think of n large, serving as a proxy for sample size, and of $\sigma > 0$ a fixed known value. If σ is unknown, one can usually replace it by a consistent estimate in the models we shall encounter in this book.

The exact meaning of dW needs further explanation. Heuristically, we may think of dW as a weak derivative of a standard Brownian motion $\{W(t) : t \in [0,1]\}$, whose existence requires a suitable notion of stochastic derivative that we do not want to develop here explicitly. Instead, we take a 'stochastic process' approach to define this stochastic differential equation, which for statistical purposes is perfectly satisfactory. Let us thus agree that 'observing the trajectory (1.5)' will simply mean that we observe a realisation of the Gaussian process defined by the application

$$g \mapsto \int_0^1 g(t)dY^{(n)}(t) \equiv \mathbb{Y}_f^{(n)}(g) \sim N\left(\langle f,g\rangle, \frac{\|g\|_2^2}{n}\right), \tag{1.6}$$

where g is any element of the Hilbert space $L^2([0,1])$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_2$. Even more explicitly, we observe all the $N(\langle f,g\rangle, \|g\|_2^2/n)$ variables, as g runs through $L^2([0,1])$. The randomness in the equation (1.5) comes entirely from the additive term dW, so after translating by $\langle f,g\rangle$ and scaling by $1/\sqrt{n}$, this means that dW is defined through the Gaussian process obtained from the action

$$g \mapsto \int_0^1 g(t)\, dW(t) \equiv \mathbb{W}(g) \sim N(0, \|g\|_2^2), \quad g \in L^2([0,1]). \tag{1.7}$$

Note that this process has a diagonal covariance in the sense that for any *finite* set of orthonormal vectors $\{e_k\} \subset L^2$ we have that the family $\{\mathbb{W}(e_k)\}$ is a multivariate standard normal variable, and as a consequence of the Kolmogorov consistency theorem (Proposition 2.1.10), $\mathbb{W}$ and $\mathbb{Y}^{(n)}$ indeed define Gaussian processes on L^2.

The fact that the model (1.5) can be interpreted as a Gaussian process indexed by L^2 means that the natural sample space $\mathcal{Y}$ in which dY from (1.5) takes values is the 'path' space $\mathbb{R}^{L^2([0,1])}$. This space may be awkward to work with in practice. In Section 6.1.1 we shall show that we can find more tractable choices for $\mathcal{Y}$ where dY concentrates with probability 1.

Gaussian Sequence Space Model

Again, to observe the stochastic process $\{\mathbb{Y}_f^{(n)}(g) : g \in L^2\}$ just means that we observe $\mathbb{Y}_f^{(n)}(g)$ for all $g \in L^2$ simultaneously. In particular, we may pick any orthonormal basis $\{e_k : k \in \mathbb{Z}\}$ of L^2, giving rise to an observation in the *Gaussian sequence space model*

$$Y_k \equiv Y_{f,k}^{(n)} = \langle f, e_k\rangle + \frac{\sigma}{\sqrt{n}}g_k, \qquad k \in \mathbb{Z}, \quad n \in \mathbb{N}, \tag{1.8}$$

where the g_k are i.i.d. of law $\mathbb{W}(e_k) \sim N(0, \|e_k\|_2^2) = N(0,1)$. Here we observe all the basis coefficients of the unknown function f with additive Gaussian noise of variance σ^2/n. Note that since the $\{e_k : k \in \mathbb{Z}\}$ realise a sequence space isometry between L^2 and the sequence space ℓ_2 of all square-summable infinite sequences through the mapping $f \mapsto \langle f, e_k \rangle$, the law of $\{Y_{f,k}^{(n)} : k \in \mathbb{Z}\}$ completely characterises the finite-dimensional distributions, and thus the law, of the process $\mathbb{Y}_f^{(n)}$. Hence, models (1.5) and (1.8) are observationally equivalent to each other, and we can prefer to work in either one of them (see also Theorem 1.2.1).

We note that the random sequence $Y = (Y_k : k \in \mathbb{Z})$ itself does not take values in ℓ_2, but we can view it as a random variable in the 'path' space $\mathbb{R}^{\ell_2}$. A more tractable, separable sample space on which $(Y_k : k \in \mathbb{Z})$ can be realised is discussed in Section 6.1.1.

A special case of the Gaussian sequence model is obtained when the space is restricted to n coefficients

$$Y_k = \theta_k + \frac{\sigma}{\sqrt{n}} g_k, \quad k = 1, \dots, n, \tag{1.9}$$

where the θ_k are equal to the $\langle f, e_k \rangle$. This is known as the *normal means model*. While itself a finite-dimensional model, it cannot be compared to the standard Gaussian linear model from the preceding section as its dimension increases as fast as n. In fact, for most parameter spaces that we will encounter in this book, the difference between model (1.9) and model (1.8) is negligible, as follows, for instance, from inspection of the proof of Theorem 1.2.1.

Multivariate Gaussian Models

To define a Gaussian white noise model for functions of several variables on $[0,1]^d$ through the preceding construction is straightforward. We simply take, for $f \in L^2([0,1]^d)$,

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \qquad t \in [0,1]^d, \quad n \in \mathbb{N}, \quad \sigma > 0, \tag{1.10}$$

where dW is defined through the action

$$g \mapsto \int_{[0,1]^d} g(t) dW(t) \equiv \mathbb{W}(g) \sim N(0, \|g\|_2^2) \tag{1.11}$$

on elements g of $L^2([0,1]^d)$, which corresponds to multivariate stochastic integrals with respect to independent Brownian motions $W_1(t_1), \dots, W_d(t_d)$. Likewise, we can reduce to a sequence space model by taking an orthonormal basis $\{e_k : k \in \mathbb{Z}^d\}$ of $L^2([0,1]^d)$.

1.2.3 Equivalence of Statistical Experiments

It is time to build a bridge between the preceding abstract models and the statistically more intuitive nonparametric fixed-design regression model (1.4). Some experience with the preceding models reveals that a statistical inference procedure in any of these models constructively suggests a procedure in the others with comparable statistical properties. Using a suitable notion of distance between statistical experiments, this intuition can be turned into a theorem, as we show in this subsection. We present results for Gaussian regression models; the general approach, however, can be developed much further to show that even highly non-Gaussian models can be, in a certain sense, asymptotically equivalent to the standard Gaussian white noise model (1.5). This gives a general justification for a

rigorous study of the Gaussian white noise model in itself. Some of the proofs in this subsection require material from subsequent chapters, but the main ideas can be grasped without difficulty.

The Le Cam Distance of Statistical Experiments

We employ a general notion of distance between statistical experiments $\mathcal{E}^{(i)}, i = 1,2$, due to Le Cam. Each experiment $\mathcal{E}^{(i)}$ consists of a sample space $\mathcal{Y}_i$ and a probability measure $P_f^{(i)}$ defined on it, indexed by a common parameter $f \in \mathcal{F}$. Let $\mathcal{T}$ be a measurable space of 'decision rules', and let

$$L : \mathcal{F} \times \mathcal{T} \to [0, \infty)$$

be a 'loss function' measuring the performance of a decision procedure $T^{(i)}(Y^{(i)}) \in \mathcal{T}$ based on observations $Y^{(i)}$ in experiment i. For instance, $T^{(i)}(Y^{(i)})$ could be an estimator for f so that $\mathcal{T} = \mathcal{F}$ and $L(f,T) = d(f,T)$, where d is some metric on $\mathcal{F}$, but other scenarios are possible. The risk under $P_f^{(i)}$ for this loss is the $P_f^{(i)}$-expectation of $L(f, T^{(i)}(Y^{(i)}))$, denoted by $R^{(i)}(f, T^{(i)}, L)$. Define also

$$|L| = \sup\{L(f,T) : f \in \mathcal{F}, T \in \mathcal{T}).$$

The *Le Cam distance* between two experiments is defined as

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \equiv \max\Big[\sup_{T^{(2)}} \inf_{T^{(1)}} \sup_{f, L:|L|=1} \left| R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L) \right|, \tag{1.12}$$

$$\sup_{T^{(1)}} \inf_{T^{(2)}} \sup_{f, L:|L|=1} \left| R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L) \right| \Big].$$

If this quantity equals zero, this means that any decision procedure $T^{(1)}$ in experiment $\mathcal{E}^{(1)}$ can be translated into a decision procedure $T^{(2)}$ in experiment $\mathcal{E}^{(2)}$, and vice versa, and that the statistical performance of these procedures in terms of the associated risk $R^{(i)}$ will be the same for any bounded loss function L. If the distance is not zero but small, then, likewise, the performance of the corresponding procedures in both experiments will differ by at most their Le Cam distance.

Some useful observations on the Le Cam distance are the following: if both experiments have a common sample space $\mathcal{Y}^{(1)} = \mathcal{Y}^{(2)} = \mathcal{Y}$ equal to a complete separable metric space, and if the probability measures $P_f^{(1)}, P_f^{(2)}$ have a common dominating measure μ on $\mathcal{Y}$, then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \le \sup_{f \in \mathcal{F}} \int_{\mathcal{Y}} \left| \frac{dP_f^{(1)}}{d\mu} - \frac{dP_f^{(2)}}{d\mu} \right| d\mu \equiv \|P^{(1)} - P^{(2)}\|_{1,\mu,\mathcal{F}}. \tag{1.13}$$

This follows from the fact that in this case we can always use the decision rule $T^{(2)}(Y)$ in experiment $\mathcal{E}^{(1)}$ and vice versa and from

$$|R^{(1)}(f,T,L) - R^{(2)}(f,T,L)| \le \int_{\mathcal{Y}} |L(f,T(Y))| |dP_f^{(1)}(Y) - dP_f^{(2)}(Y)| \le |L| \|P^{(1)} - P^{(2)}\|_{1,\mu,\mathcal{F}}.$$

The situation in which the two experiments are not defined on the sample space needs some more thought. Suppose, in the simplest case, that we can find a bi-measurable isomorphism B of $\mathcal{Y}^{(1)}$ with $\mathcal{Y}^{(2)}$, independent of f, such that

$$P_f^{(2)} = P_f^{(1)} \circ B^{-1}, \qquad P_f^{(1)} = P_f^{(2)} \circ B \quad \forall f \in \mathcal{F}.$$

Then, given observations $Y^{(2)}$ in $\mathcal{Y}^{(2)}$, we can use the decision rule $T^{(2)}(Y^{(2)}) \equiv T^{(1)}(B^{-1}(Y^{(2)}))$ in $\mathcal{E}^{(2)}$, and vice versa, and the risks $R^{(i)}$ in both experiments coincide by the image measure theorem. We can conclude in this case that

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = \Delta_{\mathcal{F}}(\mathcal{E}^{(1)},B^{-1}(\mathcal{E}^{(2)})) = 0. \tag{1.14}$$

In the absence of such a bijection, the theory of sufficient statistics can come to our aid to bound the Le Cam distance. Let again $\mathcal{Y}^{(i)}, i = 1,2$, be two sample spaces that we assume to be complete separable metric spaces. Let $\mathcal{E}^{(1)}$ be the experiment giving rise to observations $Y^{(1)}$ of law $P_f^{(1)}$ on $\mathcal{Y}^{(1)}$, and suppose that there exists a mapping $S : \mathcal{Y}^{(1)} \to \mathcal{Y}^{(2)}$ independent of f such that

$$Y^{(2)} = S(Y^{(1)}), \qquad Y^{(2)} \sim P_f^{(2)} \quad \text{on } \mathcal{Y}^{(2)}.$$

Assume, moreover, that $S(Y^{(1)})$ is a sufficient statistic for $Y^{(1)}$; that is, the conditional distribution of $Y^{(1)}$ given that we have observed $S(Y^{(1)})$ is independent of $f \in \mathcal{F}$. Then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = 0. \tag{1.15}$$

The proof of this result, which is an application of the *sufficiency principle* from statistics, is left as Exercise 1.1.

Asymptotic Equivalence for Nonparametric Gaussian Regression Models

We can now give the main result of this subsection. We shall show that the experiments

$$Y_i = f(x_i) + \varepsilon_i, \qquad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N(0,\sigma^2), \quad i = 1,\dots,n, \tag{1.16}$$

and

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t), \qquad t \in [0,1], \quad n \in \mathbb{N}, \tag{1.17}$$

are asymptotically ($n \to \infty$) equivalent in the sense of Le Cam distance. In the course of the proofs, we shall show that any of these models is also asymptotically equivalent to the sequence space model (1.8). Further models that can be shown to be equivalent to (1.17) are discussed after the proof of the following theorem.

We define classes

$$\mathcal{F}(\alpha,M) = \left\{ f : [0,1] \to \mathbb{R}, \ \sup_{x\in[0,1]} |f(x)| + \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|^\alpha} \le M \right\},$$
$$0 < \alpha \le 1, \quad 0 < M < \infty,$$

of α-Hölderian functions. Moreover, for $(x_i)_{i=1}^n$ the design points of the fixed-design regression model (1.16) and for f any bounded function defined on $[0,1]$, let $\pi_n(f)$ be the unique function that interpolates f at the x_i and that is piecewise constant on each interval $(x_{i_1}, x_i] \subset [0,1]$.

Theorem 1.2.1 *Let $(\mathcal{E}_n^{(i)} : n \in \mathbb{N}), i = 1,2,3$, equal the sequence of statistical experiments given by $i = 1$ the fixed-design nonparametric regression model (1.16); $i = 2$, the standard Gaussian white noise model (1.17); and $i = 3$, the Gaussian sequence space model (1.8),*