CHAPTER II

# The Mathematical Background

Despite the great variety of examples introduced in Chapter I, and the equally great variety of statistical questions which arise from them, we will be able to study both with just a handful of basic tools from the theory of stochastic processes: the theory of *counting processes* and their *intensity processes*, the theory of *stochastic integration*, and *martingale central limit theory*, all centering around the mathematical concepts of *martingale*, *predictable process*, and *filtration*. The present chapter surveys and summarizes the basic theory as we will need it and also gives some basic mathematical material on product-integrals and functional differentiation (the functional delta-method).

The first section gives a heuristic introduction to the basic concepts in the context of one of our basic examples: censored survival data (see Example I.3.1). Key concepts such as filtration, martingale, and predictable process are introduced informally, and their relationships in stochastic integration and martingale central limit theory are illustrated through a statistical example. The aim is to provide intuition for readers not anxious to study the general theory of stochastic processes in great depth. In fact, only quite a small part of this extensive and elaborate theory is needed.

Subsequent sections give a more formal and precise account of the theory. As we will explain later, the core material for the book is the following: basic probability theory in Sections II.2, II.3, and II.4.1, asymptotic theory in Section II.5, and product-integration in II.6. Model building and likelihood constructions need the more advanced material in the rest of Section II.4 and in Section II.7. Section II.8 is not needed at a first reading. The whole chapter is intended to form a compendium of results for later use, and the reader should not feel obliged to study it all in detail before proceeding with the rest of the book.

It would certainly be helpful if the reader already has some familiarity with such notions as *martingale* and *stopping time*, at least, in discrete time. For an introduction to discrete-time martingale theory, see almost any intermediate-level course in probability theory; e.g., Sections 7.7–7.9 of Grimmett and Stirzaker (1982); and for an extensive treatment, including the measure-theoretic approach to conditional expectation and the Radon–Nikodym derivative, see, e.g., Chung (1974) or Breiman (1968). Fleming and Harrington (1991, Chapter 1 and Appendix A) gave a useful summary.

Section II.2 sets the scene, defining stochastic processes, filtrations, and stopping times. Some important notational conventions are also introduced here; these and some others are summarized at the end of the present introductory material.

Section II.3 contains the basic theory of stochastic integration as we shall be using it, tied up with the notions of predictable process and local martingale. The continuous-time theory depends on the fundamental Doob–Meyer decomposition, separating a process into a systematic (predictable) part and a purely random (martingale) part. The elementary discrete-time version of this theorem, together with discrete-time stochastic integration, can be found, for instance, in the book by Chung (1974, Theorem 9.3.2 and Exercises 9.3.9 and 9.3.16). In discrete time, the stochastic integral, an almost trivial object, is known as Doob's "martingale transform."

Now that our basic calculus has been set out, we can start in Section II.4, studying the objects which the book is really about: counting processes. They are related to the general theory of martingales and predictable processes through the key notion of the stochastic (predictable) intensity process belonging to a given counting process. In fact, the integrated intensity process is the systematic part (or compensator) in the Doob–Meyer decomposition of the counting process itself. The basic theory of counting processes is presented in Section II.4.1. Subsequent subsections contain more specialized topics—miscellaneous results on building counting process models which will be much used in Chapter III (though less in other parts of the book).

Section II.5 introduces central limit theorems for (continuous time) martingales. The main theorem, due to Rebolledo (1979), will be continually used to prove the asymptotic normality of statistical estimators, test statistics, etc. The conditions of the main theorem are specialized to the case we will need all the time: martingales which are stochastic integrals of predictable processes with respect to counting process martingales (compensated counting processes). Here, the reader will need some familiarity with the theory of convergence in distribution (weak convergence) of stochastic processes. Fleming and Harrington (1991, Appendix B) gave a clear summary and further references.

Section II.6 contains a nonstochastic interlude on the theory of product-integrals, with its applications (in statistics and probability theory) to hazard rates and Markov processes. The theory is not especially deep: One should consider it as a way to provide an evocative notation for a simple mathe-

matical operation which crops up time and time again in apparently different contexts—likelihood expressions, building transition probabilities from transition rates, distributions from hazard rates. For a first reading, the most important thing is the definition and the applications. Material on special properties of the product-integral can be consulted when needed.

The next two sections are more specialized and could also be omitted at a first reading.

Section II.7 uses product-integrals to represent likelihoods for counting process models. The basic results are given in Section II.7.1; more advanced material (martingale connections, partial likelihood) in the other subsections are needed mainly in Chapter III for our discussion of independent and noninformative censoring. The reader must here be familiar with the notion of the Radon–Nikodym derivative of one measure with respect to another as providing a generalization of both probability density, mass function, and likelihood ratio.

Section II.8 is concerned with a very different topic. It develops the idea of differentiating functions of functions (e.g., the mapping from a distribution function to its quantile function) with a view toward an infinite-dimensional version of the delta-method (variously known as the method of propagation of errors, first-order Taylor expansion, linear approximation, etc.), familiar from elementary statistics (Rao, 1973, Section 6a.2). This supplies a calculus for generating new weak convergence results from old, extending the range of the martingale central limit theorem. This material is not central to the book and the reader could consult it when needed (mainly in Chapters IV and VIII).

Finally, Section II.9 contains historical and bibliographic remarks, together with some technical notes on specialized matters.

## *Notation*

Vectors and matrices are (usually) printed in bold type. This also applies to vector or matrix functions of time $t$. (In Chapter VIII, bold is also used for other purposes.)

Integrals are Lebesgue–Stieltjes integrals. For a crash course in modern integration theory, see, e.g., Breiman (1968, Appendix) or Rudin (1976, Chapter 11). We use both of the notations $\int \cdots dF(t)$ and $\int \cdots F(dt)$ to denote the integral of some function with respect to (the measure generated by) $F$; no difference in meaning is intended; some contexts just make one or the other notation less ambiguous or more intuitive than the other. Product-integrals ($\prod$) are defined in Section II.6.

$I(\cdots)$ is usually used for the indicator of the set $\{\cdots\}$; $\mathbf{I}$ stands for an identity matrix, $\mathcal{I}$ for an observed or asymptotic information matrix, $\imath$ for an identity mapping ($\imath(t) = t$) and $\boldsymbol{\imath}$ for a vector of identity mappings [$\boldsymbol{\imath}(t) = (t, \ldots, t)$].

When vectors really are used as vectors they are to be considered *column* vectors. Often however we just use vector notation to indicate a collection of objects, and write, e.g., $\mathbf{N} = (N_1, \ldots, N_k)$, whereby we do not distinguish between a row or column vector.

The transpose of a vector is indicated by the superscript $^\top$; $t$ usually denotes a time (also $s$, $u$, $v$, $\tau$); $\mathcal{T}$ is a collection of times (the interval $[0, \tau)$ or $[0, \tau]$, where $\tau = \infty$ is also allowed); $T$ is usually a random time (a stopping time, in fact). There is nothing special about the initial time 0, this could also have been replaced by an arbitrary value $\sigma$. The indicator function of a time interval $(s, t]$ is denoted $I_{(s,t]}$.

The fixed time interval $\mathcal{T} = [0, \tau)$ or $[0, \tau]$ will usually be in the background at any point in the book. When an integral $\int \cdots$ without limits on the integral sign is mentioned, we mean the *function* of time $t$, obtained by integrating over the interval $[0, t]$ for each $t \in \mathcal{T}$. The same convention applies to the product-integral $\prod \cdots$.

$F$ and $f$, $A$ and $\alpha$, $\Lambda$, and $\lambda$, usually denote a function and its density or derivative (also in vector or matrix versions). $\Delta F$ is the function giving the jumps of $F$, $\Delta F(t) = F(t) - F(t-)$, supposing $F$ to be right-continuous. In particular therefore, $\int \cdots F(\mathrm{d}t) = \int \cdots f(t)\,\mathrm{d}t$ if $F$ has a density $f$, whereas $\int \cdots F(\mathrm{d}t) = \sum \cdots \Delta F(t)$ if $F$ is a step-function.

$D(\mathcal{T})$ is the space of right-continuous functions with left-hand limits on $\mathcal{T}$; the so-called *cadlag* functions. The space $D(\mathcal{T})$ itself is often called the Skorohod space, reflecting the usual choice of metric on this space in classical weak convergence theory (Billingsley, 1968; Pollard, 1984).

The notations $\xrightarrow{\mathcal{D}}$ and $\xrightarrow{\mathrm{P}}$ stand for convergence in distribution and probability, respectively.

The notations $\langle \cdot \rangle$ and $[\cdot]$ for predictable and optional variation process are introduced in Section II.3. The context should always show when a time interval $[s, t]$ and when an optional covariation or "square bracket" process $[M, M']$ is meant.

## II.1. An Informal Introduction to the Basic Concepts

In subsequent sections of this chapter, we shall give a formal survey of the theory which we will use to analyze statistical models based on counting processes. By way of introduction, however, we will first discuss the key concepts in a very informal manner in the context of one of the basic examples. At the same time, we will see how the mathematical background lends itself beautifully to the treatment of statistical models formulated in terms of the *hazard rate* and possibly involving *censoring*; thus, also introducing topics treated in depth in Chapters III and IV.

Consider first a sample of $n$ (uncensored) continuously distributed survival times $X_1, \ldots, X_n$ from a survival function $S$ with hazard rate function $\alpha$; thus,

$\alpha = f/(1-F)$ where $F = 1 - S$ is the distribution function and $f$ the density of the $X_i$. The hazard rate $\alpha$ completely determines the distribution through the relations

$$S(t) = \mathrm{P}(X_i > t) = \prod_0^t [1 - \alpha(s)\,\mathrm{d}s] = \exp\left(-\int_0^t \alpha(s)\,\mathrm{d}s\right),$$

where the product-integral $\prod (1-\alpha)$ is explained later in this section, see (2.1.13), and studied in detail in Section II.6. One can interpret $\alpha$ by the heuristic

$$\mathrm{P}(X_i \in [t, t+\mathrm{d}t) | X_i \geq t) = \alpha(t)\,\mathrm{d}t. \tag{2.1.1}$$

We will consider the nonparametric estimation of the hazard rate or, rather, the cumulative or integrated hazard rate

$$A(t) = \int_0^t \alpha(s)\,\mathrm{d}s. \tag{2.1.2}$$

Typically, in survival analysis problems, complete observation of $X_1, \ldots, X_n$ is not possible. Rather, one only observes $(\tilde{X}_i, D_i)$, $i = 1, \ldots, n$, where $D_i$ is a "censoring indicator," a zero-one valued random variable describing whether $X_i$ or only a lower bound to $X_i$ is observed (really $D_i$ indicates uncensored):

$$X_i = \tilde{X}_i \quad \text{if } D_i = 1,$$
$$X_i > \tilde{X}_i \quad \text{if } D_i = 0.$$

We shall consider $\tilde{X}_1, \ldots, \tilde{X}_n$ as random times; at these times, the value of the corresponding $D_i$ becomes available, and we know whether the corresponding event is a failure or a censoring. Thus, all $n$ survival periods start together at time $t = 0$.

As an example, Figures II.1.1 and II.1.2 depict the observations of 10 randomly selected patients from the data on survival with malignant melanoma introduced in Example I.3.1: First, in the original calendar time scale, and second, in the survival time scale $t$ years since operation. This latter time scale is the one we concentrate on in our illustration of stochastic process concepts. A filled circle corresponds to $D_i = 1$ (a failure), an open circle to $D_i = 0$ (a censoring).

Further analysis is impossible without assumptions on censoring. We will make the most general assumption which still allows progress: the assumption of *independent censoring* (to which we return in Section III.2.2), which means that at any time $t$ (in the survival time scale) the survival experience in the future is not statistically altered (from what it would have been without censoring) by censoring and survival experience in the past. To formalize this notion, we must be able to talk mathematically about past and future. This will be done through the concept of a *filtration* or *history* $(\mathcal{F}_t)_{t\geq 0}$, $\mathcal{F}_t$ representing the available data at time $t$. Write $\mathcal{F}_{t-}$ correspondingly for the available data just before time $t$. A specification of $(\mathcal{F}_t)_{t\geq 0}$ can only be done

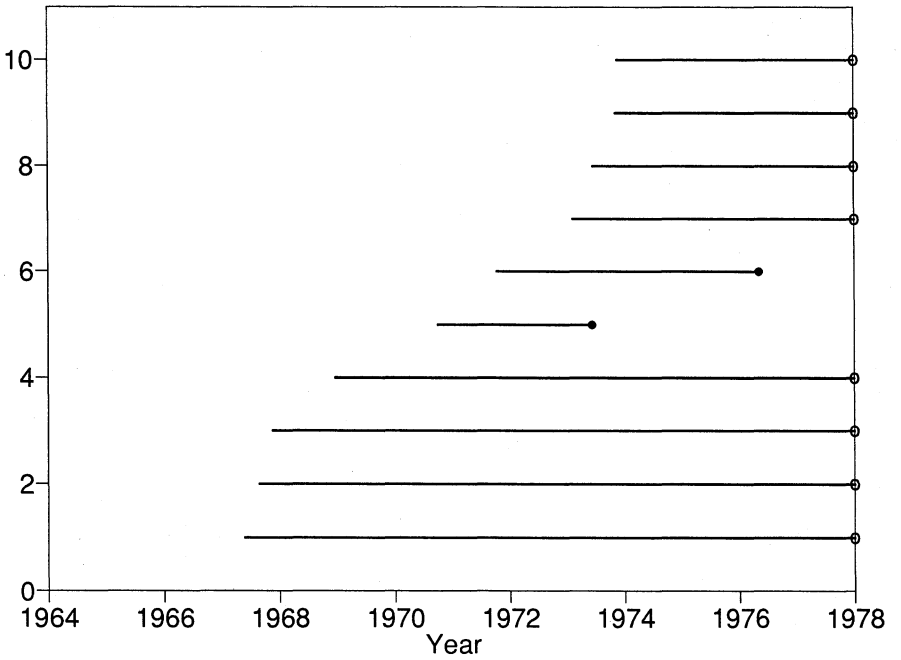


Figure II.1.1. Ten observations from the malignant melanoma study, calendar time (years).

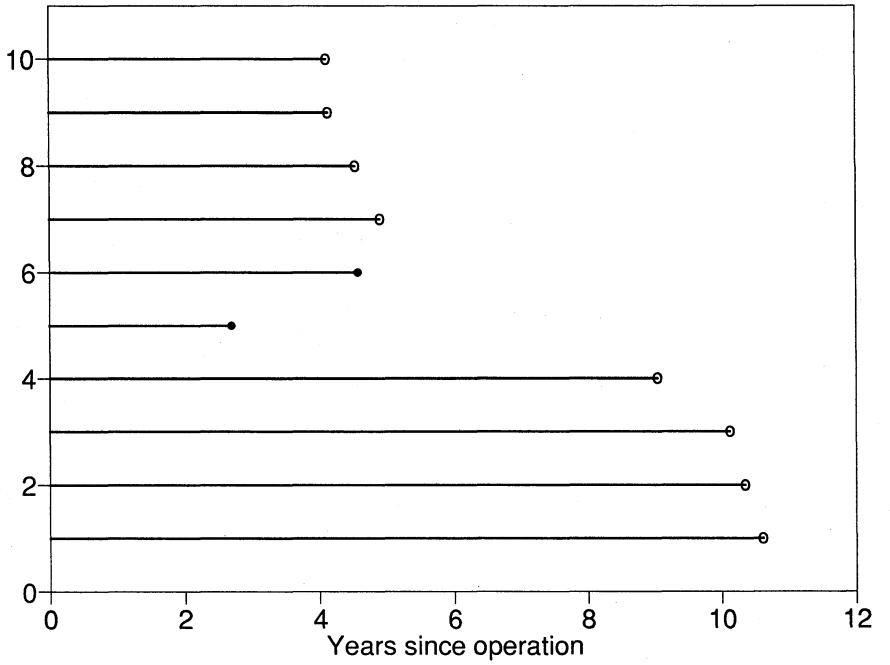


Figure II.1.2. Ten observations from the malignant melanoma study, years since operation (survival time).

relative to some observer, and different observers may collect more or less information—this interplay will be a central theme in model building; see especially Chapters III and IX. But for all obervers, as time proceeds, more information becomes available.

The notion of a filtration is defined formally in Section II.2, as an increasing family of $\sigma$-algebras defined on the sample space. In our simple example, we will simply take $\mathscr{F}_t$ to mean the values of $\tilde{X}_i$ and $D_i$ for all $i$ such that $\tilde{X}_i \le t$, otherwise just the information that $\tilde{X}_i > t$. For $\mathscr{F}_{t-}$ the obvious changes must be made: $\le$ becomes $<$ and the $>$ becomes $\ge$.

The independent censoring assumption can now be written (still very informally) as

$$\mathrm{P}(\tilde{X}_i \in [t, t+\mathrm{d}t), D_i = 1 | \mathscr{F}_{t-}) = \begin{cases} \alpha(t)\,\mathrm{d}t & \text{if } \tilde{X}_i \ge t \\ 0 & \text{if } \tilde{X}_i < t, \end{cases} \tag{2.1.3}$$

compare this to (2.1.1). Replacing the probablity on the left-hand side by the expectation of an indicator random variable, and summing over $i$, we get

$$\begin{aligned} \mathrm{E}(\#\{i\colon \tilde{X}_i \in [t, t+\mathrm{d}t), D_i = 1\} | \mathscr{F}_{t-}) &= \#\{i\colon \tilde{X}_i \ge t\} \cdot \alpha(t)\,\mathrm{d}t \\ &= Y(t)\alpha(t)\,\mathrm{d}t \\ &= \lambda(t)\,\mathrm{d}t, \end{aligned} \tag{2.1.4}$$

where we have defined the processes $Y$ and $\lambda$ by

$$Y(t) = \#\{i\colon \tilde{X}_i \ge t\},$$

the number at risk just before time $t$ for failing in the time interval $[t, t+\mathrm{d}t)$, or the size of the risk set, and

$$\lambda(t) = Y(t)\alpha(t).$$

Now formula (2.1.4) can be interpreted as a *martingale property* involving a certain *counting process*; in this case, the process $N = (N(t))_{t \ge 0}$ counting the observed failures

$$N(t) = \#\{i\colon \tilde{X}_i \le t, D_i = 1\}$$

and its *intensity process* $\lambda$. Let us write $\mathrm{d}N(t)$ or $N(\mathrm{d}t)$ for the increment $N((t+\mathrm{d}t)-) - N(t-)$ of $N$ over the small time interval $[t, t+\mathrm{d}t)$; note that this quantity is precisely the number whose conditional expectation is taken in (2.1.4). With this notation, we can, therefore, rewrite (2.1.4) as

$$\mathrm{E}(\mathrm{d}N(t) | \mathscr{F}_{t-}) = \lambda(t)\,\mathrm{d}t. \tag{2.1.5}$$

Note that the intensity process is random, through dependence on the conditioning random variables in $\mathscr{F}_{t-}$.

To explain the meaning of martingale property, first define the integrated or *cumulative intensity process* $\Lambda$ by

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s, \quad t \ge 0,$$

and the *compensated counting process* or *counting process martingale* $M$ by

$$M(t) = N(t) - \Lambda(t)$$

or, equivalently,

$$dN(t) = d\Lambda(t) + dM(t) = \lambda(t)\,dt + dM(t) = Y(t)\alpha(t)\,dt + dM(t). \quad (2.1.6)$$

Consider the conditional expectation, given the strict past $\mathcal{F}_{t-}$, of the increment (or difference) of the process $M$ over the small time interval $[t, t + dt)$; by (2.1.6), we find

$$\begin{aligned} E(dM(t)|\mathcal{F}_{t-}) &= E(dN(t) - d\Lambda(t)|\mathcal{F}_{t-}) \\ &= E(dN(t) - \lambda(t)\,dt|\mathcal{F}_{t-}) \\ &= E(dN(t)|\mathcal{F}_{t-}) - \lambda(t)\,dt \\ &= 0, \end{aligned} \quad (2.1.7)$$

where the last step is precisely the equality (2.1.5), noting that $\lambda(t)\,dt$ is fixed (nonrandom) given $\mathcal{F}_{t-}$. Now, relation (2.1.7) says that $\Lambda$ is the *compensator* of $N$, or that $M = N - \Lambda$ is a *martingale*: Such a process is characterized by the relation $E(dM(t)|\mathcal{F}_{t-}) = 0$ for all $t$.

In wide generality, we have that any counting process $N$, that is, a process taking the values 0, 1, 2, ... in turn and registering by a jump from the value $k - 1$ to $k$ the time of the $k$th occurrence of a certain type of event, has an intensity process $\lambda$ defined by $\lambda(t)\,dt = E(dN(t)|\mathcal{F}_{t-})$. The intensity process is characterized by the fact that $M = N - \Lambda$, where $\Lambda$ is the corresponding cumulative intensity process, is a martingale. Note that the processes $\lambda$ and $\Lambda$ are *predictable* (Section II.3.1); that is to say, $\lambda(t)$ is fixed *just before* time $t$, and given $\mathcal{F}_{t-}$ we know $\lambda(t)$ already [but not yet $N(t)$, for instance].

The martingale property says that the conditional expectation of increments of $M$ over small time intervals, given the past at the beginning of the interval, is zero. This is (heuristically, at least) equivalent to the more familiar definition of a martingale (see Section II.3.1)

$$E(M(t)|\mathcal{F}_s) = M(s) \quad (2.1.8)$$

for all $s < t$, which, in fact, just requires the same property for all intervals $(s, t]$: For adding up the increments of $M$ over small subintervals $[u, u + du)$ partitioning $[s + ds, t + dt) = (s, t]$, we find

$$\begin{aligned} E(M(t)|\mathcal{F}_s) - M(s) &= E(M(t) - M(s)|\mathcal{F}_s) \\ &= E\left(\int_{s<u\le t} dM(u)|\mathcal{F}_s\right) \\ &= \int_{s<u\le t} E(dM(u)|\mathcal{F}_s) \\ &= \int_{s<u\le t} E(E(dM(u)|\mathcal{F}_{u-})|\mathcal{F}_s) \\ &= 0. \end{aligned}$$

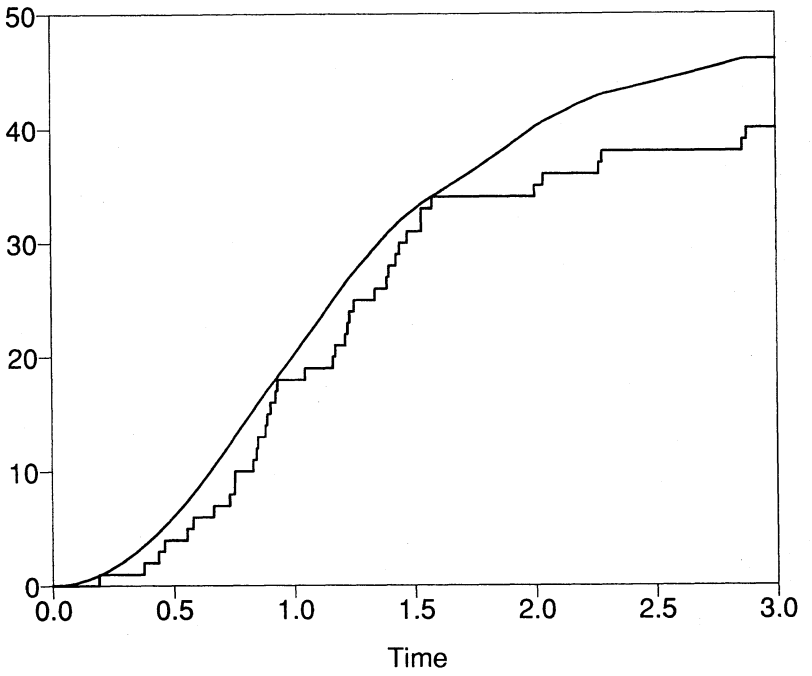


Figure II.1.3. Counting process $N(t)$ and its compensator $\Lambda(t)$, based on $n = 50$ independent randomly censored survival times with hazard $\alpha(t) = t$ (hazard for censoring distribution $0.15t$).

Version (2.1.8) of the martingale property is much easier to make the basis of a mathematical theory.

We can consider a martingale as being a pure noise process. The systematic part of a counting process is its compensator: a smoothly varying and predictable process, which, subtracted from the counting process, leaves unpredictable zero-mean noise. This is illustrated in Figures II.1.3 and II.1.4, which show a Monte Carlo realization of a counting process $N$ and its compensator $\Lambda$, and the associated martingale $M = N - \Lambda$. The counting process counts failures in a sample of 50 independent randomly censored failure times; the failure-time distribution has hazard rate $\alpha(t) = t$, the censoring distribution has hazard rate $0.15t$ (both Weibull distibutions). Censoring and failure are independent; thus, in this case $(\tilde{X}_i, D_i) = (\min(X_i, U_i), I(X_i \leq U_i))$ where the failure times $X_i$ and the censoring times $U_i$ are all independent with the specified distributions.

The notions of martingale and compensator will turn up again and again. A special case of this is the concept of the *predictable variation* $\langle M \rangle$ of a martingale (Section II.3.2). Consider the process $M^2$. Though $M$ was pure noise, $M^2$ has a tendency to increase over time. We look at its systematic component (its compensator), called $M$'s *predictable variation process* and

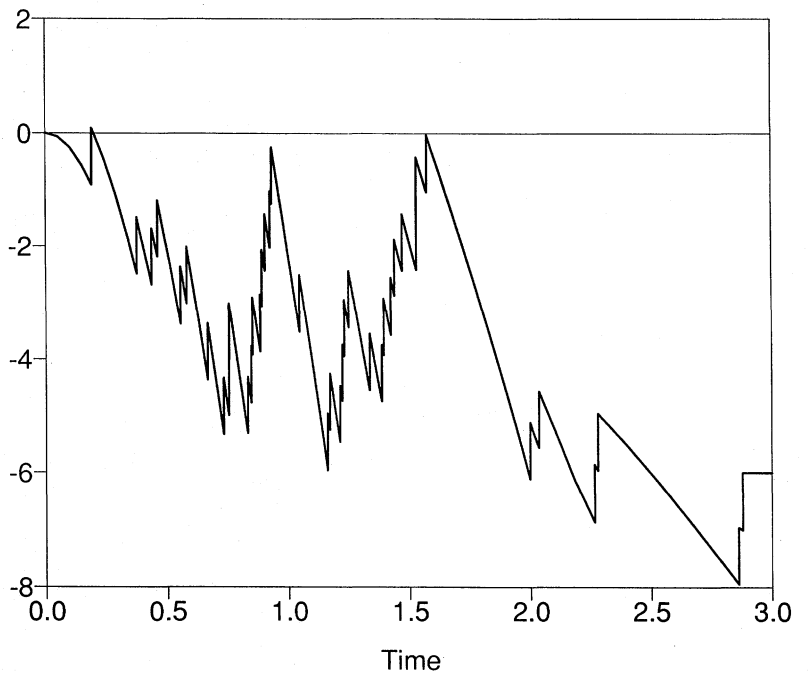


Figure II.1.4. Martingale $M(t) = N(t) - \Lambda(t)$ based on the same simulated data as in Figure II.1.3.

denoted by $\langle M \rangle$. Now

$$\begin{aligned} \mathrm{d}(M^2)(t) &= M((t + \mathrm{d}t)-)^2 - M(t-)^2 \\ &= (M(t-) + \mathrm{d}M(t))^2 - M(t-)^2 \\ &= (\mathrm{d}M(t))^2 + 2\,\mathrm{d}M(t)M(t-). \end{aligned}$$

Because $\mathrm{E}(\mathrm{d}M(t)M(t-)|\mathscr{F}_{t-}) = M(t-)\mathrm{E}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = 0$, we have

$$\mathrm{E}(\mathrm{d}(M^2)(t)|\mathscr{F}_{t-}) = \mathrm{E}((\mathrm{d}M(t))^2|\mathscr{F}_{t-}),$$

that is, the increments of the compensator of $M^2$ are the conditional *variances* of increments of $M$ (since their conditional means are zero). We summarize this very important fact as

$$\mathrm{var}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = \mathrm{d}\langle M \rangle(t).$$

Now, because no two uncensored survival times fall into the same small interval (the increments of $N$ over small time intervals are zero or one), $\mathrm{d}M(t)$ is just a zero-one random variable minus its conditional expectation $\mathrm{d}\Lambda(t)$ and, therefore, $\mathrm{var}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = \mathrm{d}\Lambda(t)(1 - \mathrm{d}\Lambda(t)) \approx \mathrm{d}\Lambda(t)$. (Note that in a discrete-time situation, the last approximation will not hold and binomial

variances will appear.) Thus, if $M$ is a compensated counting process (and the compensator in question, $\Lambda$, is continuous), then $M$'s predictable variation process $\langle M \rangle$ is simply $\Lambda$ itself.

This remarkable fact is related to the fact that for a Poisson random variable, mean and variance coincide. A counting process $N$ behaves locally at time $t$, and conditional on the past, just like a Poisson process with rate $\lambda(t)$. So, conditional means and variances of increments over small time intervals both coincide with the conditional local rate.

So far, we have only considered probabilistic matters. To introduce the next topic, stochastic integration, we shall turn to the statistical problem of nonparametric estimation of the cumulative hazard rate $A$, see (2.1.2). Using the version $dN(t) = Y(t)\alpha(t)\,dt + dM(t)$ of (2.1.6), we obtain, on multiplying throughout by $1/Y(t)$ [supposing for the moment $Y(t)$ to be positive],

$$\frac{dN(t)}{Y(t)} = \alpha(t)\,dt + \frac{dM(t)}{Y(t)}. \tag{2.1.9}$$

Now, if $dM(t)$ is just noise, the same must be true of $dM(t)$ times $1/Y(t)$: We have, because $Y(t)$ is predictable,

$$E\left(\frac{dM(t)}{Y(t)}\middle|\mathcal{F}_{t-}\right) = \frac{E(dM(t)|\mathcal{F}_{t-})}{Y(t)} = 0. \tag{2.1.10}$$

The conditional variance of the noise does change: It is now

$$\operatorname{var}\left(\frac{dM(t)}{Y(t)}\middle|\mathcal{F}_{t-}\right) = \frac{\operatorname{var}(dM(t)|\mathcal{F}_{t-})}{Y(t)^2} = d\langle M\rangle(t)\left(\frac{1}{Y(t)}\right)^2. \tag{2.1.11}$$

Let us introduce some more notation. To take account of the possibility that $Y(t)$ may be zero, define

$$J(t) = I(Y(t) > 0), \qquad H(t) = J(t)/Y(t)$$

(with $0/0 = 0$); let

$$\hat{A}(t) = \int_{0<s\le t} H(s)\,dN(s), \qquad A^*(t) = \int_{0<s\le t} J(s)\alpha(s)\,ds$$

and, finally,

$$Z(t) = \int_{0<s\le t} H(s)\,dM(s).$$

Then replacing $t$ in (2.1.9) by $s$ and integrating over $0 < s \le t$ gives

$$\hat{A}(t) = A^*(t) + Z(t), \tag{2.1.12}$$

where the left-hand side is indeed an estimator of $A(t)$: $\hat{A}(t)$ (called the Nelson–Aalen estimator and studied in Section IV.1) is just the sum over failure times up to and including time $t$ of the reciprocals of the corresponding risk set sizes; whereas, on the right-hand side, $A^*(t)$ is essentially $A(t)$ itself [we

only omit contributions $\alpha(s)\,ds$ where the risk set is empty, which hopefully hardly ever happens if nonparametric estimation of $A(t)$ is to be meaningful]. Finally, also on the right-hand side, $Z(t)$ is, by (2.1.10), the value at time $t$ of the martingale $Z$, formed on integrating the predictable process $H$ with respect to the martingale $M$. Its predictable variation process is, by (2.1.11), the integral of the *square* of $H$ with respect to the predictable variation process of $M$:

$$Z(t) = \int_{0<s\le t} H(s)\,dM(s), \qquad \langle Z\rangle(t) = \int_{0<s\le t} H(s)^2\,d\langle M\rangle(s).$$

This is a general result on *stochastic integration.* We have used it to express statistical estimation error as a martingale, from which a great deal of useful consequences can be derived, both exact or small sample results and asymptotic or large sample results. We discuss here the large sample consequences, via the *martingale central limit theorem* (Section II.5.1).

Typically, with a large sample size $n$, the random variation in the sample averages $Y(t)/n$ and $N(t)/n$ is small. Suppose, in particular, that $Y(t)/n$ for all $t$ is close to a deterministic function $y(t)$. Then we find by an easy calculation that for the martingale $W^{(n)} = \sqrt{n}(\hat{A} - A^*) = \sqrt{n}Z$ [cf. (2.1.12)], essentially $\sqrt{n}(\hat{A} - A)$, the conditional variances [by (2.1.11)] $n\lambda(t)\,dt/Y(t)^2$ of its increments $\sqrt{n}\,dM(t)/Y(t)$ are approximately equal to $\alpha(t)\,dt/y(t)$ while its many jumps are very small: of a size of the order of $1/\sqrt{n}$. So $W^{(n)}$ has an almost deterministic predictable variation process $\langle W^{(n)}\rangle \approx \int \alpha/y$, whereas its sample paths are almost continuous.

Now these properties actually characterize the (approximate) distribution of $W^{(n)}$: There is precisely one *continuous* martingale $W^{(\infty)}$ with *deterministic* predictable variation $\langle W^{(\infty)}\rangle = \int \alpha/y$, and that is a Gaussian martingale with this *variance function.* Put another way, on making the deterministic time change $t \to \langle W^{(\infty)}\rangle(t)$, $W^{(\infty)}$ becomes a standard Brownian motion or Wiener process. Plotting $\sqrt{n}(\hat{A}(t) - A(t))$ against an estimator of its predictable variation, e.g., $n\int_{0<s\le t} dN(s)/Y(s)^2$, we see for large $n$ this limiting Brownian motion: a process with independent Gaussian increments, with means zero and variances equal to the lengths of the corresponding time intervals. This result can be used, for instance, to construct confidence bands for the unknown function $A$, as we will do in Section IV.1.3.

The variance estimator $n\int_{0<s\le t} dN(s)/Y(s)^2$ is exactly equal to the sum of the squares of the increments of the martingale $W^{(n)}$ over the interval $[0, t]$. Considered as a process, this is called the *optional variation process* of $W^{(n)}$ and denoted $[W^{(n)}]$ (see Section II.3.2); thus

$$[W^{(n)}](t) = \int_{0<s\le t} \{dW^{(n)}(s)\}^2.$$

Recall that the predictable variation process $\langle W^{(n)}\rangle$ is defined by the same sum but with conditional expectations taken of each squared increment:

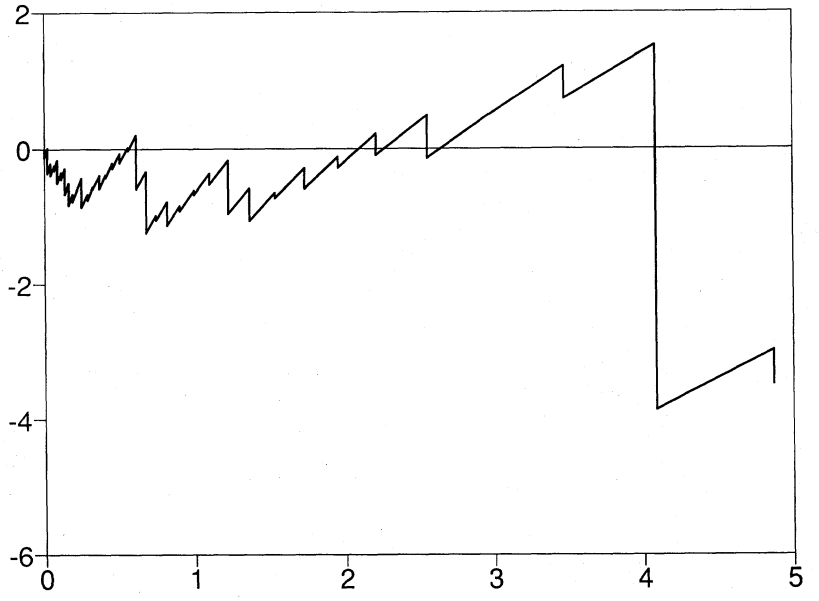


Figure II.1.5. $\sqrt{n}(\hat{A}(t) - A(t))$ plotted against its estimated variance, the optional variation process $n\int_0^t \mathrm{d}N(u)/Y(u)^2$; same data as in Figures II.1.3 and II.1.4.

$$\langle W^{(n)} \rangle(t) = \int_{0<s\leq t} \mathrm{E}(\mathrm{d}W^{(n)}(s)^2 | \mathscr{F}_{s-}).$$

For a Brownian motion, predictable and optional variations coincide and equal the (unconditional) variance function of the process.

These results are illustrated in Figure II.1.5, which shows the realization of $\sqrt{n}(\hat{A}(t) - A(t))$ plotted against its estimated variance, the optional variation process, for the same data which were used in Figures II.1.3 and II.1.4. This is beginning, indeed, to look like a typical realization of a Brownian motion.

From $A$, we can go on to estimate the survival function $S$; rewriting the relation between the two as

$$S(t) = \prod_{s=0}^{t} (1 - \mathrm{d}A(s)) \tag{2.1.13}$$

in *product-integral* notation suggests the estimator $\hat{S}(t) = \prod_{s=0}^{t}(1 - \mathrm{d}\hat{A}(s))$, a product-limit estimator; in fact, the famous Kaplan–Meier estimator (see Section IV.3). Note that for continuous $A$, because $1 - \mathrm{d}A(s) \approx \exp(-\mathrm{d}A(s))$, we can also write $S(t) = \exp(-A(t))$. However, $\hat{A}$ is a jump function and its product-integral is just a finite product and also a step-function. The statistical properties of the Kaplan–Meier estimator can be derived from slightly more elaborate martingale properties; in fact, $\hat{S}/S - 1$ turns out to be a stochastic integral with respect to $M$ and, therefore, also a martingale.

Product-integration will be explained at length in Section II.6. A formal definition of the right-hand side of (2.1.13) is as a limit of approximating finite products. Because $dA(s) = \alpha(s)\,ds$ can be interpreted as $P(X \in [s, s + ds) | X \geq s)$, see (2.1.1), we have that $1 - dA(s)$ is $P(X \geq s + ds | X \geq s)$ and multiplying such conditional probabilities over small intervals $[s, s + ds)$ partitioning $[0, t + dt)$ gives $P(X \geq t + dt)$ or just $P(X > t)$.

These estimators were derived by solving a natural estimating equation, putting the noise equal to zero in the equation $dN(t) = d\Lambda(t) + dM(t)$. In fact, a nonparametric maximum likelihood motivation is also possible, based on the fact that relation (2.1.5) essentially specifies the conditional distribution of increments of $N$ (because they can only take the values zero and one) given the past. Putting these conditional distributions together, we can build up the whole distribution of $N$.

We are neglecting here the fact that from one small time interval to the next, not just failures can occur but also other events, censorings, and their distribution has to be considered too if one wants to build up the distribution of the whole observed data. We can write, rather informally,

$$P(\text{data}) = \prod_{0<t<\infty} P(dN(t) | \mathcal{F}_{t-}) P(\text{other events in } dt | dN(t), \mathcal{F}_{t-}). \quad (2.1.14)$$

In Section II.7, we show that this likelihood expression can also be given a precise mathematical interpretation, again via the notion of product-integration.

As we mentioned, this can be used to show that the Nelson–Aalen estimator $\hat{A}$, as well as the Kaplan–Meier estimator and its generalization to Markov processes, has an interpretation as nonparametric maximum likelihood estimator; see, in particular, Section IV.1.5. The behavior of the likelihood under censoring is studied in Section III.2, and parametric maximum likelihood estimators in Chapter VI. Efficiency of tests and estimators is analyzed using large sample approximations to the likelihood in Chapter VIII.

Consider now a parametric estimation problem in which the hazard rate $\alpha$ depends on a parameter $\theta$, say. If $\theta$ does not enter in the specification of the second factors in (2.1.14), we say that we have *noninformative censoring*, to which we return in Section III.2.3. In this case, using (2.1.4) as the specification of the conditional mean of the zero-one variable $dN(t)$, we can write

$$P(\text{data}) \propto \prod_{0<t<\infty} ((\lambda^\theta(t)\,dt)^{dN(t)}(1 - \lambda^\theta(t)\,dt)^{1-dN(t)}),$$

where the intensity process, depending on $\theta$, is $\lambda^\theta(t) = Y(t)\alpha^\theta(t)$. This can be simplified somewhat. First, we can neglect the factors $dt$ in the first part of the product because these will cancel when we form likelihood ratios. Second, by a Taylor expansion, $1 - \lambda^\theta(t)\,dt \approx \exp(-\lambda^\theta(t)\,dt)$, and a product of exponentials is an exponential of a sum. This means we can write the likelihood as

$$L(\theta) \propto \left( \prod_{0<t<\infty} \lambda^\theta(t)^{\mathrm{d}N(t)} \right) \exp\left( -\int_0^\infty \lambda^\theta(t)\,\mathrm{d}t \right) \tag{2.1.15}$$

and, hence, the log-likelihood

$$\log L(\theta) = \int_0^\infty \log \lambda^\theta(t)\,\mathrm{d}N(t) - \int_0^\infty \lambda^\theta(t)\,\mathrm{d}t + \text{const}$$

and, finally, the score function

$$\begin{aligned} \frac{\partial}{\partial\theta} \log L(\theta) &= \int_0^\infty \frac{\partial}{\partial\theta} \log \lambda^\theta(t)\,\mathrm{d}N(t) - \int_0^\infty \frac{\partial}{\partial\theta} \lambda^\theta(t)\,\mathrm{d}t \\ &= \int_0^\infty \left( \frac{\partial}{\partial\theta} \log \lambda^\theta(t) \right) \mathrm{d}M^\theta(t), \end{aligned}$$

where $\mathrm{d}M^\theta(t) = \mathrm{d}N(t) - \lambda^\theta(t)\,\mathrm{d}t$.

If we had calculated the likelihood based on the data up to time $t$, we would have obtained the same results but with the integrals over $[0, \infty)$ replaced by integrals over $[0, t]$. This result shows us that the statistically extremely important score function, seen as a process (using the data up to time $t$ for each $t$), is a martingale: It can be written as a stochastic integral, this time of the derivative of the log intensity process (a predictable process) with respect to the counting process martingale. One can interpret the result as identifying the total score for the data as the sum of the scores for the infinitesimal conditional experiments: At time $t$ observe $\mathrm{d}N(t)$, a zero-one variable with mean $\lambda^\theta(t)\,\mathrm{d}t$ given $\mathscr{F}_{t-}$.

The martingale property of the score based on the data up to each time instant $t$ holds also without the assumption of noninformative censoring [which states that the second factor in (2.1.14) does not depend on $\theta$]. Without that assumption, (2.1.15) is called the *partial likelihood* for $\theta$ based on the counting process $N$ (thus ignoring censoring); see Section II.7.3. The fact that the martingale property of the *partial score process* is maintained is part of the explanation that partial-likelihood-based statistical procedures have many of the familiar asymptotic properties of ordinary likelihood methods. This will be especially apparent in Section VI.1 where maximum likelihood estimators are studied.

## II.2. Preliminaries: Processes, Filtrations, and Stopping Times

This section and the next present a compact survey of the "general theory of processes" as we need it in this book. Many topics usually central in complete treatments of this theory are ignored because in the applications in our book

they do not arise. Many purely technical points (especially concerning measurability questions) are also ignored. At some points, we diverge slightly from the usual definitions for the sake of a streamlined presentation; these divergences are discussed in the bibliographic comments. A complete but also very compact survey of the theory was given by Jacod and Shiryayev (1987) in the first chapters of their book, on which this extract is based. A survey including self-contained derivations of some of the key results was given by Fleming and Harrington (1991), who aimed at statistical applications in survival analysis. Further help in becoming at home in the vast literature surrounding this topic is given in the bibliographic comments.

We are going to model the occurrence in time of random events; in fact, discrete events occurring in continuous time. So we fix a continuous-time interval

$$\mathcal{T} = [0, \tau) \quad \text{or} \quad [0, \tau]$$

for a given terminal time $\tau$, $0 < \tau \leq \infty$. Note that the terminal time point $\tau$ may or may not be included; this varies from application to application. We write $\overline{\mathcal{T}} = [0, \tau]$, the time interval augmented with its endpoint if it was not first present.

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A *filtration*

$$(\mathcal{F}_t : t \in \mathcal{T}),$$

also called a *history*, is an *increasing right-continuous* family of sub-$\sigma$-algebras of $\mathcal{F}$. In the standard theory we use, it is often assumed also to be *complete* in the strong sense that, for every $t$, the $\sigma$-algebra $\mathcal{F}_t$ contains all P-null sets of $\mathcal{F}$. However, the assumption can be safely omitted, subject only to a very minor reformulation of the results of the standard theory; see the bibliographic comments and Jacod and Shiryayev (1987).

When the complete set of assumptions hold, we say that $(\mathcal{F}_t)$ *satisfies the usual conditions* (*les conditions habituelles*):

$$\begin{aligned} &\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F} \quad \text{for all } s < t && \text{(increasing)} \\ &\mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t \quad \text{for all } s && \text{(right continuous)} \\ &A \subset B \in \mathcal{F},\ \mathrm{P}(B) = 0 \Rightarrow A \in \mathcal{F}_0 && \text{(complete).} \end{aligned} \tag{2.2.1}$$

The $\sigma$-algebra $\mathcal{F}_t$ is interpreted as follows: It contains all events (up to null sets) whose occurrence or not is fixed by time $t$. There is also a pre-$t$ $\sigma$-algebra $\mathcal{F}_{t-}$, the smallest $\sigma$-algebra containing all $\mathcal{F}_s$, $s < t$; it contains events fixed strictly before $t$.

A stochastic process $X$ is just a time-indexed collection of random variables $(X(t) : t \in \mathcal{T})$. The process $X$ is called *adapted* (to the filtration) if $X(t)$ is $\mathcal{F}_t$ measurable for each $t$. We write $X(t, \omega)$ for the realized value of $X(t)$ at the point $\omega \in \Omega$. This way we can think of $X$ not only as a function of $\omega$ for fixed $t$ (i.e., as a random variable), but also as a function of $t$ for fixed $\omega$. This function is called a *sample path* of $X$. The process $X$ is called *cadlag* (*continu*

*à droite, limité à gauche*) if its sample paths $(X(t,\omega): t \in \mathcal{T})$, for almost all $\omega$, are right-continuous with left-hand limits. The set of cadlag functions is often denoted $D(\mathcal{T})$, the Skorohod space of weak convergence theory (Billingsley, 1968),

We often describe a filtration as being the filtration generated by a stochastic process $X$ (perhaps taking values in a quite general measurable space). This means that $\mathcal{F}_t$ is the $\sigma$-algebra generated by $X(s)$, $s \leq t$. We also have then that $\mathcal{F}_{t-}$ is generated by $X(s)$, $s < t$. Frequently, we add to $\mathcal{F}_t$ events supposed to be fixed at time zero or, equivalently, add to the random variables generating $\mathcal{F}_t$ certain other random variables whose values are supposed to be fixed at time zero.

An important result of Courrège and Priouret (1965) states that the filtration generated by a right-continuous *jump process* is right-continuous: a jump process $X$ being a process such that for each $t$ and $\omega$, $X(s,\omega)$ is constant in $s \in [t, t+\varepsilon)$ for some $\varepsilon > 0$ (depending on $t$ and $\omega$, in general). It can be shown from this that filtrations generated naturally when considering discrete events occurring in continuous time are automatically right-continuous. Extending the filtration "at time zero" as we have just described leaves it right-continuous. In particular, completion of a filtration by adding null events to every $\mathcal{F}_t$ preserves right-continuity.

A *stopping time* $T$ is a random variable taking values in $\overline{\mathcal{T}}$ such that

$$\{T \leq t\} \in \mathcal{F}_t \quad \text{for all } t \in \mathcal{T}.$$

An important example of a stopping time is the first time a process exceeds a given value: For instance, if $X$ is cadlag and adapted, then $T = \inf\{t: |X(t)| \geq c\}$ is a stopping time. Any fixed time $s$ is also a stopping time. Intuitively, the time $T$ of a random event is a stopping time if at each fixed time $t$ one can observe whether or not the event has already occurred. One can define $\sigma$-algebras $\mathcal{F}_T$ and $\mathcal{F}_{T-}$ having the interpretation as being all events which have occurred up to and including the stopping time $T$ or strictly before the stopping time. The former, $\mathcal{F}_T$, can be characterized as the $\sigma$-algebra generated by $T$ together with all the random variables $X(T)$, for any cadlag adapted process $X$. The $\sigma$-algebra of events strictly before $T$ has $X(T)$ replaced by $X(T-)$ in this description.

Many operations can be used to build new stopping times from old; for instance, the minimum or maximum of two stopping times is also a stopping time. Adding to a stopping time $T$ a non-negative $\mathcal{F}_T$ measurable random variable gives a new stopping time.

There is an intimate relation between stopping times and the main object we study in this book: counting processes. This is exemplified by considering the case of a single event occurring at a random time instant:

EXAMPLE II.2.1. The One-Jump Counting Process

Let $T$ denote the time of some random event. The indicator process $(I(T \leq t))$ is a cadlag process, equal to zero until time $T$, then jumping to the

value 1 at time $T$ (if the event ever occurs), and then staying at that value. One easily checks that the indicator process $(I(T \leq t))$ is adapted if and only if $T$ is a stopping time. This process is the simplest example of a *counting process*, to be defined formally in Section II.4.1. □

If $X$ is a stochastic process and $T$ a stopping time, it is not self-evident that $X(T)$ is, indeed, a random variable, i.e., that $X(T(\omega), \omega)$ is measurable as a function of $\omega \in \Omega$. Similar measurability questions would arise whenever we carry out operations on the process $X$ which involve its values at more than a countable number of time points $t$. However, all the processes we shall meet are nice enough that this is never a problem. For instance, all cadlag processes satisfy the measurability requirements. In the future, we will ignore this potential difficulty.

Given a stochastic process $X$ and a stopping time $T$, one often wants to consider the *stopped process* $X^T$ defined by

$$X^T(t) = X(t \wedge T),$$

where $s \wedge t = \min(s, t)$. If $X$ is cadlag and adapted and $T$ is a stopping time, then $X^T$ is cadlag and adapted too. We sketch a proof of this just to give some flavor of what is involved in a complete treatment of these basic parts of the theory. That $X^T$ is cadlag is obvious. To prove adaptedness of $X^T$, introduce $T_n = 2^{-n}(\lfloor 2^n T \rfloor + 1)$, where $\lfloor \cdot \rfloor$ denotes *entier*. Then $T_n$ is a stopping time, $T < T_n \leq T + 2^{-n}$, and $T_n \downarrow T$ as $n \to \infty$. Also, $T_n$ takes only countably many values. By an explicit calculation, one can check that $X^{T_n}$ is adapted. But for each $t$, $X^{T_n}(t) \to X^T(t)$ as $n \to \infty$ by right-continuity of $X$, showing that $X^T$ is adapted too.

Apart from arising very naturally in applications, stopping times are important in the theory through the notion of *localization*. Here a process is supposed to have some nice property only up to a stopping time, which may, however, be taken to be arbitrarily large. We then derive results for the stopped process, and afterward extend them to the original by letting the stopping times increase. In this way, nice results for a rather stringently defined class of processes may be carried over, only mildly weakened, to a much larger class.

First, we define the notion of a *localizing* sequence of stopping times. This is simply a sequence of stopping times $T_n$ which is nondecreasing and which satisfies

$$\mathrm{P}(T_n \geq t) \to 1 \quad \text{as } n \to \infty \text{ for all } t \in \mathscr{T}.$$

We then say that a process $X$ has a certain property *locally* if a localizing sequence $(T_n)$ exists such that, for each $n$, the process $I(T_n > 0)X^{T_n}$ has the property. An extremely important example of this is the following. A process $X$ is *locally bounded* if a localizing sequence $(T_n)$ and constants $c_n$ exist such that, for each $n$,

$$\sup_{t \leq T_n} |X(t)| \leq c_n \quad \text{a.s. on the event } \{T_n > 0\}.$$

Here is one special case which we will refer to often: Any left-continuous adapted process with right-hand limits is locally bounded. To see this, suppose the process $X$ has all its paths left-continuous and define $T_n = \inf\{t: |X(t)| > n\}$. One may check that this is a stopping time by first writing ($\mathbb{Q}$ the rationals)

$$\{T_n < t\} = \bigcup_{u<t, u\in\mathbb{Q}} \{|X(u)| > n\} \in \mathscr{F}_t.$$

Next, writing $\{T_n \le t\} = \bigcap_{t<s<t+\varepsilon} \{T_n < s\} \in \mathscr{F}_{t+\varepsilon}$ for any $\varepsilon > 0$ (restricting to rational $s$) and using the right-continuity of the filtration shows that $\{T_n \le t\} \in \mathscr{F}_t$. By left-continuity, the process $I(T_n > 0)X^{T_n}$ has all its paths bounded in absolute value by $n$. One may also check that the events $\{T_n \ge t\}$ are nondecreasing in $n$ and $\bigcup_n \{T_n \ge t\} = \Omega$ for each $t \in \mathscr{T}$, showing that the sequence $(T_n)$ is indeed a localizing sequence for $X$, for the property "boundedness." This example illustrates the use of adding the indicator function $I(T_n > 0)$ to the definition; otherwise, a process could only be locally bounded if its value at time zero were bounded by a constant.

*Stochastic integration*, i.e., the forming of the integral of one stochastic process with respect to another, will play a very important role in the rest of the book. This will always be for us a *pathwise* operation: for, given $\omega \in \Omega$, one forms an ordinary Lebesgue–Stieltjes integral over a given time interval. When no range of integration is explicitly given, we mean the result obtained by integrating over the interval $[0, t]$ for each $t \in \mathscr{T}$. For example, for two stochastic processes $X$ and $Y$, $\int X\,dY$ denotes the *stochastic process*

$$t \mapsto \int_0^t X(s)\,dY(s) = \int_{[0,t]} X(s)\,dY(s), \tag{2.2.2}$$

defined for each $\omega$ and $t$ such that

$$\int_{[0,t]} |X(s)||dY(s)| < \infty. \tag{2.2.3}$$

Here, $Y$ is assumed to be a cadlag process with paths of locally bounded variation; i.e., $\int_{[0,t]} |dY(s)|$ is finite for all $t \in \mathscr{T}$, for almost all $\omega \in \Omega$. We call such a process $Y$ a *finite variation process*, and the process $\int |dY|$ is called its (total) *variation process*. By convention, $Y(0-) = 0$, so that for $t = 0$, $\int_0^t X\,dY = X(0)Y(0)$. [Usually $Y(0) = 0$ in our applications.] If $X$ is nonnegative and $Y$ is nondecreasing, we need not impose the condition (2.2.3) as long as we allow the process $\int X\,dY$ to take the value $\infty$ too. A further useful notational convention is to omit the integrating function or measure $dY$ when just Lebesgue measure is meant; so $\int X$ means the process $t \mapsto \int_0^t X(s)\,ds$. We sometimes use the notation $\int_0^t X(s)Y(ds)$ instead of $\int_0^t X(s)\,dY(s)$, either to distinguish between different possible integration variables, or to emphasize the role of $Y$ as a measure rather than as a (distribution) function.

To a cadlag process $X$, we associate its left-continuous modification $X_-$, defined by

$$X_-(t) = X(t-),$$

and its jump process $\Delta X$, defined by

$$\Delta X(t) = X(t) - X(t-).$$

Two stochastic processes whose paths coincide outside of a P-null set are called *indistinguishable*. Later, when we claim uniqueness of certain processes, we mean uniqueness modulo indistinguishability. Also, when we say a process is left-continuous, right-continuous, etc., we mean that it is indistinguishable from a process all of whose paths have these properties.

# II.3. Martingale Theory

## II.3.1. Martingales, Predictable Processes, Compensators

Two kinds of stochastic processes play important and complementary roles in the general theory of processes and, hence, in the theories of stochastic integration and of counting processes: martingales and predictable processes, especially finite variation predictable processes (compensators). Martingales can be considered as a kind of pure random noise process; see the examples in Section II.1, especially Figures II.1.4 and II.1.5. On the other hand, a predictable process has a kind of regularity or semideterministic behavior. Many stochastic processes can be written as the sum of a (local) martingale and a finite variation predictable process. In such a decomposition (a generalization of the celebrated Doob–Meyer decomposition for submartingales), we are splitting the process into a random and a systematic part. The latter is called the compensator of the process because subtracted off the process, a local martingale (unsystematic noise) is left.

We give the definitions and important properties of these processes, starting with martingales and local martingales.

A *martingale* is a cadlag adapted process $M$ which is *integrable*, i.e.,

$$\mathrm{E}(|M(t)|) < \infty \quad \text{for all } t \in \mathcal{T}$$

and satisfies the *martingale property*

$$\mathrm{E}(M(t)|\mathcal{F}_s) = M(s) \quad \text{for all } s \leq t. \tag{2.3.1}$$

The process is a *submartingale* if this is replaced by the inequality

$$\mathrm{E}(M(t)|\mathcal{F}_s) \geq M(s) \quad \text{for all } s \leq t. \tag{2.3.2}$$

When we have (2.3.2) with the inequality reversed, $M$ is called a *supermartingale*. A martingale is called *square integrable* if

$$\sup_{t \in \mathcal{T}} \mathrm{E}(M(t)^2) < \infty. \tag{2.3.3}$$

A square integrable martingale can be extended to $\overline{\mathcal{T}}$; i.e., even if $\tau \notin \mathcal{T}$, the

limit $\lim_{t\to\tau} M(t)$ exists almost surely, and *defining* $M(\tau)$ as this limit, one forms a square integrable martingale on the extended time interval.

A martingale defined on $\overline{\mathscr{T}}$ is automatically uniformly integrable, in the strong sense that the set of random variables $M(T)$, for all stopping times $T$, is uniformly integrable:

$$\lim_{c\uparrow\infty} \sup_{T} \mathrm{E}(|M(T)|\, I(|M(T)| > c)) = 0.$$

A uniformly integrable martingale satisfies Doob's optional sampling theorem: that is, it satisfies the martingale property (2.3.1) with the fixed times $s \le t$ replaced by stopping times $S \le T$.

For a process to be a martingale or square integrable martingale, by definition certain integrability conditions have to be satisfied. These conditions may be hard to verify in practice or not even true at all. On the other hand, for some of the major applications (limit theory for instance) they can be avoided through the technique of localization introduced in the previous section. Especially important become the notions of a *local martingale* and a *local square integrable martingale.*

Written out in full, a *local martingale* is a process $M$ such that an increasing squence of stopping times $(T_n)$ exists,

$$\mathrm{P}(T_n \ge t) \to 1 \quad \text{as } n \to \infty \text{ for all } t \in \mathscr{T},$$

such that the stopped processes

$$I(T_n > 0)M^{T_n}$$

are martingales for each $n$. A local martingale is clearly cadlag and adapted. Not so obvious is the fact that any local martingale is actually a local uniformly integrable martingale; in other words, the localizing sequence $(T_n)$ can always be chosen so that the stopped processes are actually martingales on $\overline{\mathscr{T}}$.

Similarly, a *local square integrable martingale* is a process $M$ as above such that the localizing sequence can be chosen making $I(T_n > 0)M^{T_n}$ a square integrable martingale.

Doob's optional sampling theorem guarantees that when we consider later two different processes each having certain properties locally, e.g., a locally bounded process and a local martingale, we can choose a single localizing sequence for both processes simultaneously. This technique is used time and time again in building up the general theory of stochastic integration.

Given a local (square integrable) martingale $M$ and a particular stopping time $T$, it is often useful to know if the stopped process $I(T > 0)M^T$ is a (square integrable) martingale. Not every stopping time will have this property. We give some useful conditions for this at the end of Section II.3.2.

A class of processes complementary to the martingales is the class of *predictable processes.* A stochastic process $H$ is called predictable if, as a function

of $(t, \omega) \in \mathscr{T} \times \Omega$, it is measurable with respect to the $\sigma$-algebra on $\mathscr{T} \times \Omega$ generated by the *left-continuous adapted processes* (to be precise, generated by those adapted processes *all* of whose paths are left-continuous). Many equivalent characterizations exist; for instance, a process $H$ is predictable if and only if $H(T)$ is $\mathscr{F}_{T-}$-measurable for all stopping times $T$. Thus, the value of a predictable process at time $T$ is fixed just before the time itself. For us, it is important to note that at any rate the left-continuous adapted processes themselves are predictable (or indistinguishable from a predictable process). Also, any deterministic measurable function, considered as a stochastic process, is predictable. (This holds true because the left-continuous deterministic step-functions on $\mathscr{T}$ generate its Borel $\sigma$-algebra; considered as processes, they are trivially left-continuous and adapted. This illustrates that some right-continuous processes are predictable too. We will see more examples soon.) Occasionally useful is the fact that a cadlag predictable process is locally bounded: By predictability, one can, in effect, stop the process strictly before it exceeds a given value. (This is proved using properties of so-called predictable stopping times, which we do not introduce here.)

There is an important orthogonality between martingales and *finite variation predictable processes*. This is due to the fact that if a process at the same time is both a local martingale and a predictable finite variation process, then it is a trivial or constant process. Because continuous adapted processes are predictable (they are left-continuous and adapted), a continuous local martingale which is not constant has paths of unbounded variation on each bounded interval. For instance, standard Brownian motion (the Wiener process) on a finite time interval is a square integrable martingale with continuous paths of unbounded variation on each bounded interval.

A deep result on which most of the subsequent theory depends (compensators, stochastic integration) is the *Doob–Meyer* decomposition. First, we give the decomposition in its narrow sense (for a certain class of submartingales), and then describe how it can be extended to a much wider class of processes through the idea of localization.

Suppose $X$ is a submartingale such that the class of random variables $X(T)$, $T$ an arbitrary stopping time, is uniformly integrable; this is called a submartingale of class D (Doob). Then the *Doob–Meyer* decomposition theorem states the existence of a cadlag nondecreasing predictable process $\tilde{X}$ such that

$$M = X - \tilde{X}$$

is a uniformly integrable martingale, zero at time zero; moreover, $\tilde{X}$ is integrable [in fact, $\mathrm{E}(\tilde{X}(\tau)) < \infty$, even if $\tau \notin \mathscr{T}$]. The process $\tilde{X}$ can be constructed as a limit in probability of discrete approximations; $\tilde{X}(t)$ is approximated by $\sum \mathrm{E}(X(t_i) - X(t_{i-1}) | \mathscr{F}_{t_{i-1}})$, where $0 = t_0 < t_1 < \cdots < t_n = t$ is a fine partition of $[0, t]$. So, heuristically (cf. Section II.1),

$$\mathrm{d}\tilde{X}(t) = \mathrm{E}(\mathrm{d}X(t) | \mathscr{F}_{t-})$$

and

$$dM(t) = dX(t) - E(dX(t)|\mathcal{F}_{t-}).$$

The predictable part $\tilde{X}$ is the sum of conditional expectations, given the past, of increments of $X$ (nondecreasing because $X$ is a submartingale), whereas $M$ is the sum of increments minus their conditional expectations, hence a martingale.

The orthogonality between finite variation predictable processes and martingales gives the uniqueness of the decomposition: If we had two decompositions $X = \tilde{X} + M = X' + M'$, then $\tilde{X} - \tilde{X}' = M' - M$ would be a finite variation predictable martingale, zero at time zero, and hence constant and equal to zero throughout $\mathcal{T}$.

Obviously, for a supermartingale, one can state the same result but with $\tilde{X}$ now nonincreasing instead of nondecreasing. More generally, the difference of two submartingales of class D has a unique Doob–Meyer decomposition with the predictable part now of finite variation (the difference of two nondecreasing processes).

Next, by localization, we can avoid the integrability conditions altogether. Suppose $X$ is a cadlag adapted process. We say that $\tilde{X}$ is the *compensator* of $X$ if $\tilde{X}$ is a predictable, cadlag, and finite variation process such that $X - \tilde{X}$ is a local martingale, zero at time zero. If a compensator exists it is unique. It turns out that *$X$ has a compensator if and only if $X$ is the difference of two local submartingales.* (Part of the reason for this is that a local submartingale is locally a submartingale of class D).

Here is a simple example: the standard Poisson process.

EXAMPLE II.3.1. Poisson Process

Let $N = (N(t))$ be a standard Poisson process on the line $[0, \infty)$ with the filtration generated by the process itself (we sometimes denote this particular counting process by $\Pi$). Since the standard Poisson process has independent increments with means equal to the length of the corresponding interval, it follows that $M$ defined by $M(t) = N(t) - t$ is a martingale. Define the identity function $\imath$ by

$$\imath(t) = t.$$

The process equal to the function $\imath$ is predictable and nondecreasing, so, by uniqueness, it must be the compensator $\tilde{N}$ of $N$. In fact, $M = N - \tilde{N} = \Pi - \imath$ is locally square integrable: take as localizing sequence a sequence of constants, or the times of the jumps of the process $N$. □

Processes having a compensator are called *special semimartingales.* By definition, they can be written as the sum of a local martingale and a finite variation predictable process. A *semimartingale* (a more general object still) is the sum of a local martingale and a cadlag adapted finite variation process

(i.e., the finite variation part is not necessary predictable). Semimartingales have become a most important class of processes, but for us their general theory is not so relevant.

All (local) sub- and supermartingales have compensators. In particular, nondecreasing, non-negative locally integrable cadlag processes have compensators (which are also nondecreasing) because such processes are trivially local submartingales. By Jensen's inequality, the square of a square integrable martingale is a submartingale. Therefore, the square of a local square integrable martingale is a local submartingale and also has a nondecreasing compensator. The product of two local square integrable martingales $MM'$ can be written as a difference of two squares of local square integrable martingales $[MM' = \frac{1}{4}(M + M')^2 - \frac{1}{4}(M - M')^2]$ and has a compensator of finite variation. These special cases will turn up in the theory of the next subsection.

## II.3.2. Predictable and Optional Variation

Suppose $M$ and $M'$ are local square integrable martingales. Then, as mentioned in the previous subsection, $M^2$ is a local submartingale and $MM'$ is the difference of two local submartingales. These processes, therefore, have compensators which are denoted by $\langle M, M\rangle$ and $\langle M, M'\rangle$, respectively. As shorthand, we write $\langle M\rangle$ for $\langle M, M\rangle$. Their defining properties are therefore: $\langle M\rangle = \langle M, M\rangle$ and $\langle M, M'\rangle$ are the unique finite variation cadlag predictable processes (in fact, $\langle M\rangle$ is nondecreasing) such that

$$M^2 - \langle M\rangle$$

and

$$MM' - \langle M, M'\rangle$$

are local martingales, zero at time zero.

The process $\langle M\rangle$ is called the *predictable variation process* of $M$ and, similarly, $\langle M, M'\rangle$ the *predictable covariation process* of $M$ and $M'$. Being compensators, they can be constructed as the limits of discrete approximations allowing a heuristic interpretation (cf. Section II.1) as sums of conditional variances and covariances of increments of $M$ and $M'$:

$$\mathrm{d}\langle M\rangle(t) = \mathrm{var}(\mathrm{d}M(t)|\mathscr{F}_{t-})$$

and

$$\mathrm{d}\langle M, M'\rangle(t) = \mathrm{cov}(\mathrm{d}M(t), \mathrm{d}M'(t)|\mathscr{F}_{t-}).$$

The predictable covariation process is bilinear and symmetric, just like an ordinary covariance:

$$\begin{aligned} \langle aM + bM', M''\rangle &= a\langle M, M''\rangle + b\langle M', M''\rangle, \\ \langle M, M'\rangle &= \langle M', M\rangle. \end{aligned} \tag{2.3.4}$$

The predictable (co)variation process is also called the *pointed brackets process of* $M$ (and $M'$). If

$$\langle M, M' \rangle = 0,$$

we say that $M$ and $M'$ are *orthogonal*. This means that the product $MM'$ is itself a local martingale.

We often deal with a vector of local square integrable martingales $\mathbf{M} = (M_1, \ldots, M_k)$. We then write $\langle \mathbf{M} \rangle = (\langle M_h, M_j \rangle)$ for the *matrix* of predictable covariation processes of each pair of components of $\mathbf{M}$. The matrix of processes $\langle \mathbf{M}, \mathbf{M}' \rangle$ is defined similarly.

The predictable variation process, at time $t$, of a local square integrable martingale $M$ is the limit in probability of approximations of the form $\sum \operatorname{var}(M(t_i) - M(t_{i-1}) | \mathcal{F}_{t_{i-1}})$ for ever finer partitions $0 = t_0 < t_1 \cdots < t_n = t$. When we omit the conditional expectation here but just take the limit (in probability) of sums of squares $\sum (M(t_i) - M(t_{i-1}))^2$, we obtain another nondecreasing process called the *optional variation process* of $M$. This new process exists when $M$ is just a local martingale (not necessarily locally square integrable). It is not predictable, in general. It is denoted by $[M]$ and also called $M$'s *square brackets process*, as well as $M$'s quadratic variation process. Analogously, we have an *optional covariation process* or square bracket process (or quadratic covariation process) $[M, M']$ for two local martingales $M$ and $M'$; and we can define a matrix optional covariation process for two vector local martingales. When $M$ and $M'$ are continuous, their optional covariation process coincides with the predictable covariation process. On the other hand, when $M$ and $M'$ are finite variation local martingales (as will always be the case in our applications), the optional variation process has an explicit and simple form

$$[M](t) = \sum_{s \le t} \Delta M(s)^2 = M(t)^2 - 2 \int_0^t M(s-)\, \mathrm{d}M(s),$$

$$\begin{aligned}[M, M'](t) &= \sum_{s \le t} \Delta M(s) \Delta M'(s) \\ &= M(t)M'(t) - \int_0^t M(s-)\, \mathrm{d}M'(s) - \int_0^t M'(s-)\, \mathrm{d}M(s).\end{aligned}$$

The process $[M]$ is nondecreasing, cadlag, and

$$M^2 - [M]$$

is a local martingale. If $[M]$ is actually locally integrable, then the local martingale $M$ is actually locally square integrable and $\langle M \rangle$ is the compensator of $[M]$. More generally, $\langle M, M' \rangle$ is the compensator of $[M, M']$. From this, it follows that if $M$ and $M'$ are finite variation local square integrable martingales with no jump times in common, so $\Delta M \Delta M' = 0$ and hence $[M, M'] = 0$, then $\langle M, M' \rangle = 0$ and $M$ and $M'$ are orthogonal (their product is a local martingale).

We exhibit the predictable and optional variation processes for the martingale $M$ of Example II.3.1, the compensated Poisson process, and also for the Wiener process or standard Brownian motion.

EXAMPLE II.3.2. Poisson Process (Continued)

Let $N = \Pi$ be a standard Poisson process as in Example II.3.1 and define $M = \Pi - \iota$, where $\iota$ is the identity function $\iota(t) = t$. Using the fact that the variance of a Poisson random variable equals its mean and the independent increments of the Poisson process, one can check that $M^2 - \iota$ is a martingale; so, by uniqueness, the process $\iota$ is the compensator of $M^2$, that is, $\langle M \rangle(t) = t$ for all $t$, as well as being the compensator of $\Pi$ itself. On the other hand, because $M$ is a finite variation local martingale with jumps of size $+1$ only at the jump times of $\Pi$, we find that $[M] = \Pi$. □

EXAMPLE II.3.3. Standard Brownian Motion (the Wiener Process)

It can be shown that there exists a *continuous* process $W$ which is zero at time zero and has independent normally distributed increments with means zero and variances equal to the lengths of the corresponding intervals. With respect to the natural filtration generated by the process, $W$ is a local square integrable martingale (localized by constant times) and $\langle W \rangle = [W] = \text{var}(W) = \iota$, the identity function. $W$ is called *the Wiener process* or *standard Brownian motion.* More generally, a time-transformed Brownian motion or Wiener process has continuous sample paths, independent, normally distributed mean zero increments. Its predictable and optional variation process are both equal to its variance function, a continuous nondecreasing function. □

The pointed and square bracket processes can be used to verify if a given stopping time localizes a local square integrable martingale. We may omit the multiplicative factor $I(T > 0)$, superfluous anyway if the local martingale is zero at time zero. For a local square integrable martingale $M$ and a stopping time $T$, we have that $M^T$ is a square integrable martingale if and only if

$$\mathrm{E}\langle M \rangle(T) < \infty$$

and also if and only if

$$\mathrm{E}[M](T) < \infty.$$

If $M$ is a finite variation local martingale and $T$ a stopping time, then $M^T$ is a uniformly integrable martingale of integrable variation if and only if

$$\mathrm{E} \int_0^T |\mathrm{d}M(s)| < \infty.$$

Finally, we will make much use of the following simple way to check that a stopping time localizes a local martingale. If $M$ is the difference of two

cadlag increasing processes $X$ and $\tilde{X}$, it suffices to verify that either $EX(T) < \infty$ or $E\tilde{X}(T) < \infty$ (by the use of the optional stopping theorem, localizing sequences of stopping times, and monotone convergence). For instance, applying this to $[M] - \langle M \rangle$, it follows that $\langle M \rangle$ has a finite expectation at a given time (or stopping time) if and only if $[M]$ does.

## II.3.3. Stochastic Integration

We shall only be concerned with stochastic integrals which have a pathwise interpretation, as we stated in Section II.2: $\int X \, dY$ is the ordinary pathwise Lebesgue–Stieltjes integral (over $[0, t]$, for each $t \in \mathcal{T}$) defined for processes $X$ and $Y$ such that $\int_0^t |X(s,\omega)||Y(ds,\omega)|$ is finite for almost all $\omega$, for each $t \in \mathcal{T}$. However, this integral has special and valuable properties when the integrand $X$ is a *predictable process* $H$ and we integrate with respect to a process $Y$ which is a *local (square integrable) martingale* $M$. Under the appropriate (local) integrability conditions, the resulting process $\int H \, dM$ is a local (square integrable) martingale. Moreover, its predictable and optional variation processes can be obtained simply from those of $M$.

**Theorem II.3.1.** *Suppose $M$ is a finite variation local square integrable martingale, $H$ is a predictable process, and $\int H^2 \, d[M]$ is locally integrable or $\int H^2 \, d\langle M \rangle$ is just locally finite (automatically true if $H$ is locally bounded). Then $\int H \, dM$ is a local square integrable martingale, and*

$$\left[\int H \, dM\right] = \int H^2 \, d[M],$$

$$\left\langle \int H \, dM \right\rangle = \int H^2 \, d\langle M \rangle.$$

The predictable process $H$ is locally bounded if it is left-continuous and has right-hand limits.

With weaker assumptions than those of Theorem II.3.1, we get weaker conclusions. Suppose always that $M$ is a finite variation local martingale and $H$ is predictable. If we just assume the process $\int |H||dM|$ is locally integrable (automatically true if $H$ is locally bounded), then $\int H \, dM$ is a local martingale and $[\int H \, dM] = \int H^2 \, d[M]$. Alternatively, suppose the process $\int H^2 \, d[M]$ is locally integrable (automatically true if $H$ is locally bounded and $M$ is locally square integrable). Then $\int H \, dM$ is a local square integrable martingale and $[\int H \, dM] = \int H^2 \, d[M]$.

Formulas for predictable and optional covariation processes of stochastic integrals follow the same form: We have

$$\left\langle \int H \, dM, \int K \, dM' \right\rangle = \int HK \, d\langle M, M' \rangle, \tag{2.3.5}$$

$$\left[\int H\,dM, \int K\,dM'\right] = \int HK\,d[M, M']. \tag{2.3.6}$$

Sometimes we use vector and matrix versions of these formulas; for instance, for vectors $\mathbf{M}$ and $\mathbf{M}'$ and matrices $\mathbf{H}$ and $\mathbf{K}$, we have

$$\left\langle \int \mathbf{H}\,d\mathbf{M}, \int \mathbf{K}\,d\mathbf{M}' \right\rangle = \int \mathbf{H}\,d\langle \mathbf{M}, \mathbf{M}' \rangle \mathbf{K}^{\top}, \tag{2.3.7}$$

$$\left[\int \mathbf{H}\,d\mathbf{M}, \int \mathbf{K}\,d\mathbf{M}'\right] = \int \mathbf{H}\,d[\mathbf{M}, \mathbf{M}']\mathbf{K}^{\top}. \tag{2.3.8}$$

The following remarks give some indication of how these results can be derived. To begin with, consider as integrand a predictable process $H$ of the form $H(t) = XI_{(u,v]}(t)$, where $X$ is an $\mathcal{F}_u$-measurable random variable and $u < v$ are two fixed time points. This process is adapted and left-continuous, and hence predictable. Suppose $M$ is a martingale and $\mathrm{E}(\int_0^\tau |\mathrm{H}||dM|) = \mathrm{E}(|X| \int_{(u,v]} |dM|) < \infty$. Then it is easy to check from the definitions that $\int H\,dM$ is a martingale and $[\int H\,dM] = \int H^2\,d[M]$. If $M$ is square integrable and $\mathrm{E}(\int_0^\tau H^2\,d\langle M \rangle) = \mathrm{E}(X^2(\langle M \rangle(v) - \langle M \rangle(u))) < \infty$, one can also easily check that $\int H\,dM$ is a square integrable martingale with $\langle \int H\,dM \rangle = \int H^2\,d\langle M \rangle$. Next, the results are extended to predictable processes which are a sum of a finite number of such processes; these are called *simple* predictable processes. Now we note that the class of predictable processes is generated by the simple predictable processes and apply monotone class arguments and localization to get general results.

# II.4. Counting Processes

## II.4.1. Counting Processes and Their Intensities

A multivariate counting process is a stochastic process which can be thought of as registering the occurrences in time of a number of types of disjoint, discrete events. We suppose that either a filtration is already given, relative to which the process is adapted, or that one constructs the so-called self-exciting filtration generated by the process itself; see the following text.

So, suppose a filtration $(\mathcal{F}_t: t \in \mathcal{T})$ on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is given, satisfying the usual conditions (2.2.1), except possibly completeness. A multivariate counting process

$$\mathbf{N} = (N_1, \ldots, N_k)$$

is a vector of $k$ adapted cadlag processes, all zero at time zero, with paths which are piecewise constant and nondecreasing, having jumps of size $+1$ only; no two components jumping simultaneously. We suppose $N_h(t)$ is al-

most surely finite for each $h$ and all $t \in \mathcal{T}$. If $\tau \notin \mathcal{T}$, then $N_h(\tau) = \lim_{t \uparrow \tau} N_h(t)$ may, however, be infinite. The process is called *self-exciting* if the accompanying filtration is that generated by the process itself; the filtration $(\mathcal{N}_t)$ defined by

$$\mathcal{N}_t = \sigma(\mathbf{N}(s): s \leq t) \quad \text{for all } t,$$

possibly augmented by null sets. Often, we deal with a larger filtration $(\mathcal{F}_t)$ having the special form

$$\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{N}_t;$$

thus, $\mathcal{F}_t$ is generated by $\mathcal{F}_0$ together with $\{\mathbf{N}(s), s \leq t\}$. These filtrations are automatically right-continuous and increasing and can be made complete as well if desired (see Section II.2).

To a counting process $\mathbf{N}$ we can associate a sequence of *jump times* and *jump marks*, giving the time and type of each event. Because no two components of $\mathbf{N}$ jump simultaneously, the sum of the components is also a counting process. Defining then

$$N_\cdot = \sum_{h=1}^{k} N_h,$$

we can construct stopping times

$$0 < T_1 \leq T_2 \leq T_3 \leq \cdots$$

and random variables

$$J_1, J_2, \ldots,$$

taking values in $\{1, 2, \ldots, k\} \cup \{0\}$ such that for $n \leq N_\cdot(\tau)$, we have $T_n \in \mathcal{T}$, $J_n \neq 0$, $T_n > T_{n-1}$,

$$N_\cdot(T_n) = n \quad \text{and} \quad \Delta N_{J_n}(T_n) = 1.$$

For $n > N_\cdot(\tau)$, we set $J_n = 0$ and $T_n = \tau$ just to keep everything well defined. As just said, the variables $T_n$ are stopping times; also $J_n$ is $\mathcal{F}_{T_n}$-measurable. This is essentially the notation for a *marked point process*, to which we return at the end of this subsection.

Because the components of a counting process $\mathbf{N}$ are adapted, cadlag, locally bounded ($0 \leq N_h^{T_n} \leq n$), and nondecreasing, they are local submartingales and have compensators $\Lambda_h$ which we collect together in a vector $\mathbf{\Lambda}$. Thus, each $\Lambda_h$ is a nondecreasing predictable process, zero at time zero, such that

$$M_h = N_h - \Lambda_h \tag{2.4.1}$$

is a local martingale. Because $0 \leq \Delta N_h \leq 1$, the same can be shown to hold for $\Delta \Lambda_h$ for each $h$. [Consider, for instance, the predictable process $I(\Delta \Lambda_h > 1)$ and look at its stochastic integral with respect to $M$. The result should be a local martingale. However, $N_h$ makes jumps of size one only, so

$\int I(\Delta\Lambda_h > 1)(dN_h - d\Lambda_h)$ is nonincreasing and zero at time zero, and must, therefore, be identically zero.] The compensator $\Lambda_h$ can be considered as the *integrated* or *cumulative* stochastic intensity of the counting process $N_h$, for reasons which will become clear later [cf. (2.4.4)–(2.4.7)]. We define $\Lambda_\cdot = \sum_{h=1}^{k} \Lambda_h$; $\Lambda_\cdot$ is the compensator of the counting process $N_\cdot$, so we also have $0 \leq \Delta\Lambda_\cdot \leq 1$.

We next show that $M_h$ is a local square integrable martingale, with a predictable variation process (the compensator of $M_h^2$) which can be simply described in terms of $\Lambda_h$ (the compensator of $N_h$). Now the sequence of stopping times $T_n$ can be considered as a localizing sequence, making $\mathbf{N}$ locally bounded because $N_\cdot(T_n) \leq n$ for each $n$. Also, because $\Lambda_h$ is cadlag and predictable, it is locally bounded. Combining localizing times for $N_h$ and $\Lambda_h$, one finds that, for each $h$, $M_h = N_h - \Lambda_h$ is a *local square integrable martingale*, in fact, of locally bounded variation. A little further argument shows that $M_h$ may be localized to a square integrable martingale by the sequence $T_n$ [or by any other sequence for which $EN_\cdot(T_n) < \infty$ for each $n$].

As we mentioned, the predictable variation process of $\mathbf{M}$ can be simply expressed in terms of $\mathbf{\Lambda}$. One finds that

$$\begin{aligned} \langle M_h \rangle &= \Lambda_h - \int \Delta\Lambda_h \, d\Lambda_h, \\ \langle M_h, M_{h'} \rangle &= -\int \Delta\Lambda_h \, d\Lambda_{h'} \quad (h \neq h'). \end{aligned} \tag{2.4.2}$$

In particular, when $\mathbf{\Lambda}$ is *continuous*, this becomes

$$\begin{aligned} \langle M_h \rangle &= \Lambda_h, \\ \langle M_h, M_{h'} \rangle &= 0 \quad (h \neq h'). \end{aligned} \tag{2.4.3}$$

This means that the counting process martingales are *orthogonal* (cf. Section II.3.2) when an intensity process exists.

These facts can be proved by first calculating $[\mathbf{M}]$, the matrix of processes $[M_h, M_{h'}]$, and showing that its compensator $\langle \mathbf{M} \rangle$ has the elements just described. The continuous case is especially simple, so we look at that first. In that case, $M_h$ is a local finite variation martingale with jumps coinciding with those of $N_h$ and of the same size ($+1$ only). So the sum of squares of jumps of $M_h$ equals $N_h$:

$$[M_h] = N_h.$$

It is locally integrable with compensator $\Lambda_h$, so $M_h$ is locally square integrable and $\langle M_h \rangle = \Lambda_h$. On the other hand, for $h \neq h'$, $M_h$ and $M_{h'}$ have no jumps in common at all, so $[M_h, M_{h'}] = 0$. The compensator of this process is the zero process, so $\langle M_h, M_{h'} \rangle = 0$ too. The same arguments were used in the special case of the standard Poisson process; see Examples II.3.1 and II.3.2.

In general, when $\mathbf{\Lambda}$ may have jumps, direct calculation gives

$$[M_h] = \sum (\Delta M_h)^2 = \int \Delta M_h \, dM_h$$

$$= \int (\Delta N_h - \Delta \Lambda_h)(dN_h - d\Lambda_h)$$

$$= N_h - 2 \int \Delta \Lambda_h \, dN_h + \int \Delta \Lambda_h \, d\Lambda_h,$$

where we use the fact that $(\Delta N_h)^2 = \Delta N_h$. We want to calculate the compensator of this process. The compensator of the first of the three terms on the right-hand side is already known to be $\Lambda_h$. For the second, note that $\Delta \Lambda_h$ is a bounded predictable process and $M_h$ a local martingale, so $\int \Delta \Lambda_h \, dM_h = \int \Delta \Lambda_h \, dN_h - \int \Delta \Lambda_h \, d\Lambda_h$ is a local martingale too. Thus, the predictable process $\int \Delta \Lambda_h \, d\Lambda_h$ is the compensator of $\int \Delta \Lambda_h \, dN_h$. The third term is already predictable and, therefore, its own compensator. Combining, we find that $\Lambda_h - \int \Delta \Lambda_h \, d\Lambda_h$ is the compensator of $[M_h]$. This gives us the required result for $\langle M_h \rangle$. For the predictable covariation, a similar calculation (for $h \neq h'$) gives

$$[M_h, M_{h'}] = -\int \Delta \Lambda_h \, dM_{h'} - \int \Delta \Lambda_{h'} \, dM_h + \int \Delta \Lambda_h \, d\Lambda_{h'}.$$

The compensator of this process is

$$-\int \Delta \Lambda_h \, d\Lambda_{h'} - \int \Delta \Lambda_{h'} \, d\Lambda_h + \int \Delta \Lambda_h \, d\Lambda_{h'} = -\int \Delta \Lambda_h \, d\Lambda_{h'}.$$

In most of this book, we will be interested in the so-called *absolutely continuous case*, the case in which *intensities* exist. We say that $N_h$ has *intensity process* $\lambda_h$ if $\lambda_h$ is a predictable process and

$$\Lambda_h(t) = \int_0^t \lambda_h(s) \, ds \quad \text{for all } t \tag{2.4.4}$$

(some authors would not require predictability of $\lambda_h$). If $N_h$ is integrable and $\lambda_h$ is left-continuous with right-hand limits, one can easily show that

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{E}(N_h(t + \Delta t) - N_h(t) | \mathscr{F}_t) = \lambda_h(t+) \quad \text{a.s.} \tag{2.4.5}$$

Under further conditions, one can strengthen this to

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(N_h(t + \Delta t) - N_h(t) \geq 1 | \mathscr{F}_t) = \lambda_h(t+) \quad \text{a.s.} \tag{2.4.6}$$

or even to

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(N_h(t + \Delta t) - N_h(t) = 1 | \mathscr{F}_t) = \lambda_h(t+) \quad \text{a.s.}, \tag{2.4.7}$$

justifying the name we have given to $\lambda_h$. Whereas (2.4.5)–(2.4.7) are useful in the interpretation of the intensity process, the regularity conditions needed when taking them as starting point (see Aven, 1985) do not have any further role, and none of (2.4.5)–(2.4.7) have direct operational value, so there is no use in demanding them to hold.

Clearly, an integrated or cumulative intensity process $\Lambda$ exists and is unique, up to indistinguishability, by the existence and uniqueness of compensator. An intensity process $\lambda$, if it exists, is, however, not unique because we have only defined it here by requiring predictability and $\int \lambda = \Lambda$. (If $\Lambda$ is absolutely continuous and its derivative satisfies a modest measurability condition, then a predictable version does exist.) If an intensity process which is left-continuous with right-hand limits *exists*, then it will be unique, among such processes, and up to indistinguishability.

We next give one of the most important examples of a counting process. Though it is extremely simple, it is the basis of a great deal of further developments. This is the case of a one-component one-jump self-exciting counting process (cf. Example II.2.1).

EXAMPLE II.4.1. One-Jump Process

Let $T$ be a non-negative random variable with absolutely continuous distribution function $F$, survival function $S = 1 - F$, density $f$, and hazard rate $\alpha = f/S$. Let $\tau_F$ be the upper limit of the support of $F$, but take $\tau = \infty$ and $\mathscr{T} = [0, \infty]$ even if $\tau_F < \infty$. Note that

$$\int_0^t \alpha(u)\,du = -\log(1 - F(t)) < \infty$$

for all $t < \tau_F$ though $\int_0^{\tau_F} \alpha = \infty$. Define the univariate counting process $N$ by

$$N(t) = I(T \leq t)$$

and let $(\mathscr{N}_t)$ be the filtration it generates; in fact,

$$\mathscr{N}_t = \sigma\{N(s)\colon s \leq t\} = \sigma\{T \wedge t, I(T \leq t)\}.$$

Define the left-continuous adapted process $Y$ by

$$Y(t) = I(T \geq t).$$

We show that $N$ has compensator $\Lambda$ defined by

$$\Lambda(t) = \int_0^t Y(s)\alpha(s)\,ds$$

and, hence, $N$ has intensity process $\lambda$ defined by

$$\lambda(t) = Y(t)\alpha(t).$$

To do this, we note that $\Lambda$ is predictable (it is continuous and adapted), so we need only verify that $M = N - \Lambda$ is a local martingale. In fact, it is a square

integrable martingale on $\overline{\mathcal{F}}$ by our general theory earlier. To show the martingale property, it suffices to verify that

$$\mathrm{E}(M(\infty)|\mathcal{N}_t) = M(t)$$

because this implies, for $s < t$,

$$\begin{aligned}\mathrm{E}(M(t)|\mathcal{N}_s) &= \mathrm{E}(\mathrm{E}(M(\infty)|\mathcal{N}_t)|\mathcal{N}_s)\\ &= \mathrm{E}(M(\infty)|\mathcal{N}_s)\\ &= M(s).\end{aligned}$$

Consider, first, the case $t = 0$. Because $\mathcal{N}_0$ is trivial, we must show $\mathrm{E}M(\infty) = 0$. Now $N(\infty) = 1$, whereas

$$\begin{aligned}\mathrm{E}\Lambda(\infty) &= \mathrm{E}\int_0^{\tau_F} Y(t)\alpha(t)\,\mathrm{d}t\\ &= \int_0^{\tau_F} \mathrm{P}(T \geq t)\alpha(t)\,\mathrm{d}t\\ &= \int_0^{\tau_F} S(t)\frac{f(t)}{S(t)}\mathrm{d}t = \int_0^{\tau_F} f(t)\,\mathrm{d}t = 1.\end{aligned}$$

We next show that the result for general $t$ follows from that for $t = 0$. When conditioning on $\mathcal{N}_t = \sigma(T \wedge t, I(T \leq t))$ we have to consider two separate cases: conditional on $T = s \leq t$ for some $s \leq t$ and conditional on $T > t$. In the first case, $M(\infty) = M(t) = M(s)$ and there is nothing to prove. In the second case, $M(\infty) - M(t) = 1 - \int_t^{\tau_F} I(T \geq s)\alpha(s)\,\mathrm{d}s$ and we must show that the expectation of this random variable, given $T > t$, is zero. But conditional on $T > t$, $T$ now has hazard rate $\alpha I_{(t,\tau_F)}$. So this case is just the same as the case $t = 0$ only with a different hazard rate. □

Various extensions of this example are almost as easy to obtain. For instance, if $T$ does not have a density, then one may still define an *integrated* or *cumulative hazard function* $A$ by $A(t) = \int_0^t (1 - F(s-))^{-1}\,\mathrm{d}F(s)$. The compensator of $N$ is the process $\Lambda = \int Y\,\mathrm{d}A$ with $Y$ defined as above. The calculation above goes through essentially unchanged. We see that, in this example, $N$ has an *intensity process* if and only if the distribution of $T$ has a density.

Another extension to both counting process and filtration is to let, at time $T$, further information become available in the form of a *mark* $J$, a random variable taking values in $\{1,\ldots,k\}$, telling us the *type* of the event which has just occurred. In fact, this puts us in the so-called *competing risks models*; see Example III.1.5. Define a $k$-variate counting process $\mathbf{N}$ by letting $N_h(t) = I(T \leq t, J = h)$, $t \in [0,\infty)$, $h \in \{1,\ldots,k\}$. Let the process be self-exciting by taking $\mathcal{N}_t = \sigma\{T \wedge t, I(T \leq t), J \cdot I(T \leq t)\}$. Suppose $T$ has survival function $S$ and that $(T, J)$ has a joint density $f_h(t)$ with respect to Lebesgue times counting measure. Then with $\alpha_h = f_h/S$, one finds that the compensator

of $N_h$ is $\Lambda_h = \int Y\alpha_h$, where $Y(t) = I(T \geq t)$ as before; so $N_h$ has intensity process $\lambda_h(t) = \alpha_h(t)Y(t)$. The calculation of Example II.4.1 needs only minor modifications.

With these extensions, a completely general result for the intensity of a self-exciting counting process can be finally obtained. We will give this result (Jacod's representation) in Section II.7.1.

Earlier we have shown how, for the counting process martingale $\mathbf{M} = \mathbf{N} - \mathbf{\Lambda}$, the predictable and optional covariation process $\langle \mathbf{M} \rangle$ and $[\mathbf{M}]$ can be expressed in terms of $\mathbf{\Lambda}$ and $\mathbf{N}$. These relations have consequences via Theorem II.3.1 for integrals of predictable processes with respect to $\mathbf{M}$. Suppose, for simplicity, that the counting process has an intensity process $\boldsymbol{\lambda}$; so $\mathbf{\Lambda} = \int \boldsymbol{\lambda}$ is absolutely continuous. Let $\mathbf{H}$ be a $p \times k$ matrix of locally bounded predictable processes; for instance, $\mathbf{H}$ might be left-continuous and adapted; or cadlag and predictable; or deterministic and locally bounded. Let $\int \mathbf{H}\,\mathrm{d}\mathbf{M}$ denote the $p$-vector of rowwise sums of integrals of components of $\mathbf{H}$ with respect to the $k$ components of $\mathbf{M}$. By Theorem II.3.1, the result is a vector of finite variation local square integrable martingales. Using diag of a vector to denote the diagonal matrix with the components of the vector on the diagonal, we obtain the following proposition [cf. (2.3.7) and (2.3.8)]:

**Proposition II.4.1.** *Let the counting process* $\mathbf{N}$ *have intensity process* $\boldsymbol{\lambda}$, *let* $\mathbf{M} = \mathbf{N} - \int \boldsymbol{\lambda}$, *and let* $\mathbf{H}$ *be locally bounded and predictable. Then* $\mathbf{M}$ *and* $\int \mathbf{H}\,\mathrm{d}\mathbf{M}$ *are local square integrable martingales with*

$$\begin{aligned} \langle \mathbf{M} \rangle &= \operatorname{diag} \int \boldsymbol{\lambda}, \\ [\mathbf{M}] &= \operatorname{diag} \mathbf{N}, \\ \left\langle \int \mathbf{H}\,\mathrm{d}\mathbf{M} \right\rangle &= \int \mathbf{H} \operatorname{diag} \boldsymbol{\lambda} \mathbf{H}^{\top}, \\ \left[ \int \mathbf{H}\,\mathrm{d}\mathbf{M} \right] &= \int \mathbf{H}\,\mathrm{d}(\operatorname{diag} \mathbf{N}) \mathbf{H}^{\top}. \end{aligned} \tag{2.4.8}$$

Written out componentwise, and with $\delta_{hl}$ denoting a Kronecker delta, these important equations become

$$\begin{aligned} \langle M_h, M_l \rangle &= \delta_{hl} \int \lambda_h, \\ [M_h, M_l] &= \delta_{hl} N_h, \\ \left\langle \sum_h \int H_{jh}\,\mathrm{d}M_h, \sum_l \int H_{j'l}\,\mathrm{d}M_l \right\rangle &= \sum_h \int H_{jh} H_{j'h} \lambda_h, \\ \left[ \sum_h \int H_{jh}\,\mathrm{d}M_h, \sum_l \int H_{j'l}\,\mathrm{d}M_l \right] &= \sum_h \int H_{jh} H_{j'h}\,\mathrm{d}N_h. \end{aligned} \tag{2.4.9}$$

### II.4.1.1. *Marked Point Processes*

Sometimes more than a finite number of different types of events are possible though the process counting all of them is still a counting process (there are a finite total number of events in finite time intervals). The different types may even vary continuously; think, for instance, of a continuous numerical measurement becoming available at the time of each possible event. It is now no longer convenient to count each type separately. Rather, one counts over aggregates of types; for instance, measurements in certain intervals or, more generally, in any given Borel set. The processes counting events with type in some sets will each be counting processes, but there can be overlap for overlapping sets.

We will accommodate these possibilities by considering $\mathbf{N}$ and $\mathbf{\Lambda}$ as *measures* on the product space: time cross type (or mark). The counting process $\mathbf{N}$ corresponds to a counting measure on the set of points $(T_1, J_1)$, $(T_2, J_2)$, etc. (jump time, jump type, or mark). Usually one replaces $\mathbf{N}$ and $\mathbf{\Lambda}$ by $\mu$ and $\nu$ but we will continue to use the old symbols with the new interpretation.

So let $(E, \mathscr{E})$ be some measurable space of marks or types. In the case of an ordinary multivariate counting process, $E$ is just $\{1, \ldots, k\}$ and $\mathscr{E}$ is its power set: all possible subsets of $E$. We say that $\mathbf{N}$ is a marked point process with respect to a given filtration and a given mark space if $\mathbf{N}$ is a (random) counting measure on the product space $(\mathscr{T} \times E, \mathscr{B}(\mathscr{T}) \otimes \mathscr{E})$, where $\mathscr{B}$ denotes a Borel $\sigma$-algebra and where $\mathbf{N}$ satisfies the following conditions: $N_A$ defined by

$$N_A(t) = \mathbf{N}([0, t] \times A)$$

is a counting process for every $A \in \mathscr{E}$ (an adapted cadlag step-function, jumps of size $+1$ only, zero at time zero). It follows from this that, for disjoint sets $A_1, \ldots, A_k \in \mathscr{E}$, the process $(N_{A_1}, \ldots, N_{A_k})$ is a multivariate counting process in the ordinary sense.

The process $(N_{A_1}, \ldots, N_{A_k})$, therefore, has a compensator $(\Lambda_{A_1}, \ldots, \Lambda_{A_k})$. It turns out that one can extract all these compensators from a single so-called *predictable measure* $\mathbf{\Lambda}$ on $(\mathscr{T} \times E, \mathscr{B}(\mathscr{T}) \otimes \mathscr{E})$ in the same way as is done for $\mathbf{N}$:

$$\Lambda_A(t) = \mathbf{\Lambda}([0, t] \times A).$$

Letting $(T_n, J_n)$ be the points of the marked point process, it turns out that $T_n$ is a stopping time and $J_n$ is $\mathscr{F}_{T_n}$-measurable [with values in $(E, \mathscr{E})$] for each $n = 1, 2, \ldots$.

## II.4.2. The Innovation Theorem

In this and the next subsections, we present a collection of results showing how intensity processes change under various operations: change of filtra-

tion, formation of products, starting, stopping and filtering. This provides the model-builder with a wide range of techniques for deriving new counting process models from old as we do in Chapter III. Another technique (change of probability measure) will be discussed in Section II.7.1, as well as how to start from scratch.

Consider a counting process $\mathbf{N}$, adapted to both of two filtrations $(\mathcal{F}_t)$ and $(\mathcal{G}_t)$ with $\mathcal{F}_t \subseteq \mathcal{G}_t$ for all $t$; we say the filtrations are *nested*. This corresponds to two different levels of information, or two observers. In both cases, $\mathbf{N}$ is observed. Suppose the counting process has intensity process $\lambda$ with respect to the larger filtration $(\mathcal{G}_t)$. One may ask what its intensity process with respect to the smaller filtration $(\mathcal{F}_t)$ is. This will generally differ from $\lambda$; after all, one is given less information to condition on. The answer is given by the *innovation theorem*, which states that there exists an $(\mathcal{F}_t)$-predictable process $\tilde{\lambda}$ such that

$$\tilde{\lambda}(t) = \mathrm{E}(\lambda(t)|\mathcal{F}_{t-}) \quad \text{for all } t$$

and such that $\tilde{\lambda}$ is the $(\mathcal{F}_t)$-intensity process of $\mathbf{N}$. (More generally, $t$ can be replaced by any $(\mathcal{F}_t)$-*predictable stopping time* $T$.)

If $\mathbf{N}$ has intensity process $\lambda$ with respect to a filtration $(\mathcal{G}_t)$ and $\mathbf{N}$ is adapted and $\lambda$ is predictable with respect to a smaller filtration $(\mathcal{F}_t)$, then $\lambda$ is also the $(\mathcal{F}_t)$-intensity process of $\mathbf{N}$, by uniqueness of the intensity process.

## II.4.3. Product Spaces

Probability models are often built by combining independent components in a product space. This construction is also valuable in building models based on counting processes and proceeds as follows.

Suppose we are given two filtered probability spaces $(\Omega^{(i)}, \mathcal{F}^{(i)}, (\mathcal{F}_t^{(i)}, t \in \mathcal{T}), \mathrm{P}^{(i)})$, $i = 1, 2$, on each of which is defined a counting process $\mathbf{N}^{(i)}$ with intensity process $\lambda^{(i)}$. We can now form the product space on which the processes $\mathbf{N}^{(i)}$ and $\lambda^{(i)}$ are defined in the obvious way, and naturally define a product filtration by

$$\Omega = \Omega^{(1)} \times \Omega^{(2)}, \qquad \mathcal{F} = \mathcal{F}^{(1)} \otimes \mathcal{F}^{(2)},$$

$$\mathcal{F}_t = \mathcal{F}_t^{(1)} \otimes \mathcal{F}_t^{(2)}, \qquad \mathrm{P} = \mathrm{P}^{(1)} \otimes \mathrm{P}^{(2)}.$$

Actually, $(\mathcal{F}_t)$ is not necessarily right-continuous. A sufficient condition for this is that $\mathcal{F}_t^{(i)} = \mathcal{F}_0^{(i)} \vee \sigma\{\mathbf{N}^{(i)}(s): s \leq t\}$ in which case

$$\mathcal{F}_t = \mathcal{F}_0^{(1)} \otimes \mathcal{F}_0^{(2)} \vee \sigma\{\mathbf{N}^{(1)}(s), \mathbf{N}^{(2)}(s): s \leq t\}$$

which is obviously right-continuous. It is easy to check that the processes $\lambda_h^{(i)}(t)$ are also predictable, viewed as defined on $\Omega$ and with respect to P and $(\mathcal{F}_t)$. Moreover, the $N_h^{(i)} - \int \lambda_h^{(i)}$ remain local martingales on the product space. Thus, the intensity processes of $\mathbf{N}^{(i)}$ remain the same and, indeed,

$(\mathbf{N}^{(1)}, \mathbf{N}^{(2)})$ forms a multivariate counting process with intensity process $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$ because the probability is zero of having simultaneous jumps of $\mathbf{N}^{(1)}$ and $\mathbf{N}^{(2)}$ (by continuity of their compensators).

Instead of this combination of independent component processes, we will occasionally need to combine *conditionally independent* components. The situation now is that there is one probability space $(\Omega, \mathscr{F}, \mathrm{P})$ on which $\mathbf{N}(t) = (\mathbf{N}^{(1)}(t), \mathbf{N}^{(2)}(t))$ is defined; we consider two filtrations $(\mathscr{F}_t^{(1)})$ and $(\mathscr{F}_t^{(2)})$ and assume that they are *conditionally independent* given some $\sigma$-algebra $\mathscr{A} \subseteq \mathscr{F}$, that is, if $A \in \mathscr{F}_t^{(1)}$, $B \in \mathscr{F}_t^{(2)}$, $C \in \mathscr{A}$, $\mathrm{P}(C) > 0$, then

$$\mathrm{P}(A \cap B | C) = \mathrm{P}(A|C)\mathrm{P}(B|C).$$

Often, $\mathscr{F}_t^{(i)} = \sigma\{\mathbf{N}^{(i)}(s): s \leq t\}$. The $\sigma$-algebra $\mathscr{A}$ models events which are supposed to happen "at time 0."

Suppose that the counting processes $\mathbf{N}^{(1)}$, $\mathbf{N}^{(2)}$ have intensity processes $\boldsymbol{\lambda}^{(1)}$, $\boldsymbol{\lambda}^{(2)}$ with respect to the two filtrations $(\mathscr{F}_t^{(1)})$ and $(\mathscr{F}_t^{(2)})$. Define

$$\mathscr{F}_t = \mathscr{A} \vee \mathscr{F}_t^{(1)} \vee \mathscr{F}_t^{(2)},$$

so that if the $\mathscr{F}_0^{(i)}$ are trivial, we have $\mathscr{F}_0 = \mathscr{A}$. One may check directly that any $(\mathscr{A} \vee \mathscr{F}_t^{(1)})$-martingale is also an $(\mathscr{F}_t)$-martingale, which is the key step in verifying that the $(\mathscr{F}_t)$-compensator of $(\mathbf{N}^{(1)}, \mathbf{N}^{(2)})$ may be obtained by combining the $(\mathscr{A} \vee \mathscr{F}_t^{(i)})$-compensators of $\mathbf{N}^{(i)}(t)$. This leads to the desired result that $(\mathbf{N}^{(1)}, \mathbf{N}^{(2)})$ has intensity process $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$ with respect to the combined filtration $(\mathscr{F}_t)$.

## II.4.4. Starting, Stopping, and Filtering

That *stopping* a counting process preserves its intensity (or compensator) is a nice exercise in stochastic integration: Let $\mathbf{N}$ be a multivariate counting process with compensator $\boldsymbol{\Lambda}$, let $\mathbf{M} = \mathbf{N} - \boldsymbol{\Lambda}$, and let $T$ be a stopping time. The indicator function $I_{[0,T]}$ of the stochastic interval $[0, T]$ is left-continuous and adapted, hence predictable and (locally) bounded. Observe that $\mathbf{N}^T = \int I_{[0,T]} \mathrm{d}\mathbf{N}$ and $\boldsymbol{\Lambda}^T = \int I_{[0,T]} \mathrm{d}\boldsymbol{\Lambda}$ (componentwise integration). Because $\mathbf{M}^T = \mathbf{N}^T - \boldsymbol{\Lambda}^T = \int I_{[0,T]} \mathrm{d}\mathbf{M}$, we find that $\mathbf{M}^T$ is a vector of local martingales. Also $\mathbf{N}^T$ is a new multivariate counting process, counting events up to and including time $T$ only. Because the integral of one predictable process with respect to another is again predictable, $\boldsymbol{\Lambda}^T$ is a finite variation predictable process and is, therefore, by uniqueness of compensators, the compensator of $\mathbf{N}^T$. If $\boldsymbol{\Lambda}$ admits of an intensity $\boldsymbol{\lambda}$, we see that $\mathbf{N}^T$ has the intensity process $\boldsymbol{\lambda} I_{[0,T]}$.

One now can proceed to reduce the filtration replacing $(\mathscr{F}_t)$ by $(\mathscr{F}_{t \wedge T})$. Because $\mathbf{N}^T$ and $\boldsymbol{\Lambda}^T$ are adapted with respect to this smaller filtration, by the innovation theorem (Section II.4.2), $\boldsymbol{\lambda} I_{[0,T]}$ remains the intensity process of $\mathbf{N}^T$ with respect to it.

More generally, one could consider $\tilde{\mathbf{N}} = \int C \, d\mathbf{N}$ for any predictable, zero-one valued process $C$; $\tilde{\mathbf{N}}$ is a counting process with intensity $\lambda C$. We call this *filtering*, which we discuss at length in Section III.4.

We now discuss how *starting* a counting process (at a stopping time), while conditioning on the past at that moment, also preserves its intensity. Let

$$\mathbf{N} = (N_h, h = 1, \ldots, k)$$

be a multivariate counting process on a space $(\Omega, \mathcal{F})$ with compensator $\Lambda$ and intensity process $\lambda$ with respect to a filtration $(\mathcal{F}_t)$. Furthermore, we let $V$ be an $(\mathcal{F}_t)$-stopping time and consider an event $A \in \mathcal{F}_V$. The process $\mathbf{N}$ started at $V$ is defined as

$$_V\mathbf{N}(t) = \mathbf{N}(t) - \mathbf{N}(t \wedge V).$$

We want to study the process $\mathbf{N}$, starting from the time $V$, *given* that the event $A$ (prior to $V$) has actually occurred. We call the process $_V\mathbf{N}$, *under this conditional distribution*, a left-truncated process. The proposition below states that left-truncation of $\mathbf{N}$ by the event $A$ (before $V$) preserves the intensity of $\mathbf{N}$ after time $V$. For ease of presentation, we suppose $\mathrm{P}(A) > 0$.

**Proposition II.4.2.** *The left-truncated counting process* $_V\mathbf{N}$ *has intensity process* $_V\lambda = \lambda I_{(V,\infty)}$, *i.e.*,

$$_V\lambda(t) = \lambda(t) I_{(V,\infty)}(t),$$

*with respect to the filtration* $_V\mathcal{F}_t$ *given by*

$$_V\mathcal{F}_t = \mathcal{F}_t \vee \mathcal{F}_V$$

*and the conditional probability* $\mathrm{P}^A$ *given by*

$$\mathrm{P}^A(F) = \mathrm{P}(F \cap A)/\mathrm{P}(A), \quad F \in \mathcal{F}.$$

A proof of Proposition II.4.2 was given by Andersen et al. (1988). A result similar to Proposition II.4.2 can be obtained for *any* event $A \in \mathcal{F}_V$, not necessarily satisfying $\mathrm{P}(A) > 0$, using the technical apparatus of proper regular conditional probabilities and Blackwell spaces; see Jacobsen (1982, Exercise 8, p. 51, and Appendix 1.)

# II.5. Limit Theory

## II.5.1. A Martingale Central Limit Theorem

Our main tool for studing the large sample properties of statistical methods for counting process models will be the martingale central limit theorem. There exist many versions of this, as well as generalizations to semi-

martingales. Here we present a version close to that given by Rebolledo (1980a) for locally square integrable martingales because in our applications we will be applying it to stochastic integrals with respect to the basic counting process martingales, which are locally square integrable. As we suggested in Section II.1, two conditions are required for a local square integrable martingale to be approximately Gaussian: It must have close to continuous paths, and its predictable variation process must be close to deterministic.

For each $n = 1, 2, \ldots$, let $\mathbf{M}^{(n)} = (M_1^{(n)}, \ldots, M_k^{(n)})$ be a vector of $k$ local square integrable martingales, possibly defined on different sample spaces and filtrations for each $n$. Also, for each $\varepsilon > 0$, let $\mathbf{M}_\varepsilon^{(n)}$ be a vector of $k$ local square integrable martingales, containing all the jumps of components of $\mathbf{M}^{(n)}$ larger in absolute value than $\varepsilon$: So $M_h^{(n)} - M_{\varepsilon,h}^{(n)}$ is also a local square integrable martingale and $|\Delta M_h^{(n)} - \Delta M_{\varepsilon,h}^{(n)}| \leq \varepsilon$. We write $\langle \mathbf{M}^{(n)} \rangle$ for the $k \times k$ matrix of processes $\langle M_h^{(n)}, M_{h'}^{(n)} \rangle$ (Section II.3.2).

Next, let $\mathbf{M}^{(\infty)}$ be a continuous Gaussian vector martingale with $\langle \mathbf{M}^{(\infty)} \rangle = [\mathbf{M}^{(\infty)}] = \mathbf{V}$, a continuous deterministic $k \times k$ positive semidefinite matrix-valued function on $\mathscr{T}$, with positive semidefinite increments, zero at time zero. So $\mathbf{M}^{(\infty)}(t) - \mathbf{M}^{(\infty)}(s) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}(t) - \mathbf{V}(s))$ (a multivariate normal distribution) and is independent of $(\mathbf{M}^{(\infty)}(u); u \leq s)$ for all $0 \leq s \leq t$. Given a function $\mathbf{V}$ with the above-mentioned properties, the process $\mathbf{M}^{(\infty)}$ always exists.

With these preparations made, we can now state the martingale central limit theorem in a convenient form.

**Theorem II.5.1** (Rebolledo's theorem). *Let $\mathscr{T}_0 \subseteq \mathscr{T}$ and consider the conditions*

$$\langle \mathbf{M}^{(n)} \rangle(t) \xrightarrow{\mathrm{P}} \mathbf{V}(t) \quad \text{for all } t \in \mathscr{T}_0 \text{ as } n \to \infty, \tag{2.5.1}$$

$$[\mathbf{M}^{(n)}](t) \xrightarrow{\mathrm{P}} \mathbf{V}(t) \quad \text{for all } t \in \mathscr{T}_0 \text{ as } n \to \infty, \tag{2.5.2}$$

$$\langle M_{\varepsilon h}^{(n)} \rangle(t) \xrightarrow{\mathrm{P}} 0 \quad \text{for all } t \in \mathscr{T}_0, h \text{ and } \varepsilon > 0 \text{ as } n \to \infty. \tag{2.5.3}$$

*Then either of* (2.5.1) *and* (2.5.2), *together with* (2.5.3), *imply*

$$(\mathbf{M}^{(n)}(t_1), \ldots, \mathbf{M}^{(n)}(t_l)) \xrightarrow{\mathscr{D}} (\mathbf{M}^{(\infty)}(t_1), \ldots, \mathbf{M}^{(\infty)}(t_l)) \quad \text{as } n \to \infty \tag{2.5.4}$$

*for all $t_1, \ldots, t_l \in \mathscr{T}_0$; moreover, both* (2.5.1) *and* (2.5.2) *then hold.*

*If, furthermore, $\mathscr{T}_0$ is dense in $\mathscr{T}$ and contains $\tau$ if $\tau \in \mathscr{T}$, then the same conditions imply*

$$\mathbf{M}^{(n)} \xrightarrow{\mathscr{D}} \mathbf{M}^{(\infty)} \quad \text{in } (D(\mathscr{T}))^k \text{ as } n \to \infty \tag{2.5.5}$$

*and $\langle \mathbf{M}^{(n)} \rangle$ and $[\mathbf{M}^{(n)}]$ converge uniformly on compact subsets of $\mathscr{T}$, in probability, to $\mathbf{V}$.*

Here, $(D(\mathscr{T}))^k$ is the space of $\mathbb{R}^k$-valued cadlag functions on $\mathscr{T}$ endowed with the Skorohod topology, and $\xrightarrow{\mathscr{D}}$ denotes weak convergence as described, e.g., by Billingsley (1968). Condition (2.5.1) states that the predictable variation

processes converge in probability to a deterministic function; the *Lindeberg condition* (2.5.3) states that the jumps of $\mathbf{M}^{(n)}$ become small as $n \to \infty$.

We reformulate these conditions in the case of most interest to us, stochastic integrals with respect to counting processes having a continuous compensator. For each $n = 1, 2, \ldots$, let $\mathbf{N}^{(n)}$ be a $k_n$-variate counting process with intensity process $\boldsymbol{\lambda}^{(n)}$. Let $\mathbf{H}^{(n)}$ be a $k \times k_n$ matrix of locally bounded predictable processes. Now define

$$M_j^{(n)}(t) = \sum_{h=1}^{k_n} \int_0^t H_{jh}^{(n)}(s)(\mathrm{d}N_h^{(n)}(s) - \lambda_h^{(n)}(s)\,\mathrm{d}s),$$

$$M_{j\varepsilon}^{(n)}(t) = \sum_{h=1}^{k_n} \int_0^t H_{jh}^{(n)}(s) I(|H_{jh}^{(n)}(s)| > \varepsilon)(\mathrm{d}N_h^{(n)}(s) - \lambda_h^{(n)}(s)\,\mathrm{d}s),$$

for $j = 1, \ldots, k$ and $\varepsilon > 0$; $\mathbf{M}^{(n)} = (M_1^{(n)}, \ldots, M_k^{(n)})$ and $\mathbf{M}_\varepsilon^{(n)} = (M_{1\varepsilon}^{(n)}, \ldots, M_{k\varepsilon}^{(n)})$. Then $\mathbf{M}_\varepsilon^{(n)}$ contains all the jumps of $\mathbf{M}^{(n)}$ larger than $\varepsilon$ and both are $k$-variate local square integrable martingales. The bracket processes in (2.5.1)–(2.5.3) are given (cf. Proposition II.4.1) by

$$\langle M_j^{(n)}, M_{j'}^{(n)} \rangle(t) = \sum_{h=1}^{k_n} \int_0^t H_{jh}^{(n)}(s) H_{j'h}^{(n)}(s) \lambda_h^{(n)}(s)\,\mathrm{d}s, \tag{2.5.6}$$

$$[M_j^{(n)}, M_{j'}^{(n)}](t) = \sum_{h=1}^{k_n} \int_0^t H_{jh}^{(n)}(s) H_{j'h}^{(n)}(s)\,\mathrm{d}N_h^{(n)}(s), \tag{2.5.7}$$

$$\langle M_{j\varepsilon}^{(n)}, M_{j\varepsilon}^{(n)} \rangle(t) = \sum_{h=1}^{k_n} \int_0^t (H_{jh}^{(n)}(s))^2 I(|H_{jh}^{(n)}(s)| > \varepsilon) \lambda_h^{(n)}(s)\,\mathrm{d}s. \tag{2.5.8}$$

We will usually deal with situations in which $k_n$ is fixed and $\lambda_h^{(n)}(s)$ gets larger as $n \to \infty$ for all $h$ and $s$. To balance this, we will arrange matters so that $H_{jh}^{(n)}(s)$ gets smaller [taking care of (2.5.3)] and so that $H_{jh}^{(n)}(s) H_{j'h}^{(n)}(s) \lambda_h^{(n)}(s)$ converges in probability to a finite, nonzero deterministic limit as $n \to \infty$, for each $j, j'$, and $h$ [taking care of (2.5.1)]; see Section II.1 for a simple example. Thus, the following situation will frequently arise: For some sequence of processes $X^{(n)}$, it is given that

$$X^{(n)}(s) \xrightarrow{\mathrm{P}} f(s) \quad \text{as } n \to \infty \tag{2.5.9}$$

for almost all $s \in \mathscr{T}$, where the deterministic function $f$ satisfies

$$\int_0^\tau |f(s)|\,\mathrm{d}s < \infty. \tag{2.5.10}$$

Under which supplementary conditions may we conclude

$$\int_0^\tau X^{(n)}(s)\,\mathrm{d}s \xrightarrow{\mathrm{P}} \int_0^\tau f(s)\,\mathrm{d}s? \tag{2.5.11}$$

If $\tau < \infty$, then a sufficient condition for (2.5.11) is obviously $\|X^{(n)} - f\|_\infty \xrightarrow{\mathrm{P}} 0$, where $\|\cdot\|_\infty$ denotes the supremum norm on $\mathscr{T}$. However, this is often too

strong or too difficult to verify. What we really need is a kind of dominated convergence theorem giving convergence of integrals under assumptions of pointwise convergence of the integrands. Two weaker sets of conditions, one requiring uniform integrability (and leading to convergence in mean) and the other requiring uniform, in probability, bounds (and leading to convergence in probability) are now given:

**Proposition II.5.2** (Helland, 1983). *Suppose* (2.5.9) *holds and, furthermore,*

$$\lim_{C\uparrow\infty} \sup_n \mathrm{E}(|X^{(n)}(s)| I(|X^{(n)}(s)| > C)) = 0 \quad \textit{for all } s \tag{2.5.12}$$

*(i.e.,* $(X^{(n)}(s), n = 1, 2, \ldots)$ *is uniformly integrable) and*

$$\mathrm{E}|X^{(n)}(s)| \leq k(s) \quad \textit{for all } s, n, \tag{2.5.13}$$

*where*

$$\int_0^\tau k(s)\,\mathrm{d}s < \infty.$$

*Then* (2.5.10) *also holds and*

$$\mathrm{E}\left(\sup_t \left| \int_0^t X^{(n)}(s)\,\mathrm{d}s - \int_0^t f(s)\,\mathrm{d}s \right|\right) \to 0 \quad \textit{as } n \to \infty. \tag{2.5.14}$$

**Proposition II.5.3** (Gill, 1983b). *Suppose* (2.5.9) *and* (2.5.10) *hold and, furthermore, for all* $\delta > 0$, *there exists* $k_\delta$ *with* $\int_0^\tau k_\delta < \infty$ *such that*

$$\liminf_{n\to\infty} \mathrm{P}(|X^{(n)}(s)| \leq k_\delta(s) \textit{ for all } s) \geq 1 - \delta. \tag{2.5.15}$$

*Then*

$$\sup_t \left| \int_0^t X^{(n)}(s)\,\mathrm{d}s - \int_0^t f(s)\,\mathrm{d}s \right| \xrightarrow{\mathrm{P}} 0. \tag{2.5.16}$$

Under both sets of conditions, we get (2.5.11) and more. The two sets are overlapping (neither implies the other) though Helland's conditions (called by him "convergence boundedly in $L_1$") are typically stronger than Gill's.

Proposition II.5.3 has the following very simple proof. If we replace $X^{(n)}$ by the process $\mathrm{sign}(X^{(n)})(|X^{(n)}| \wedge k_\delta)$, we obtain a sequence of bounded processes $X_\delta^{(n)}$, converging pointwise in probability to the function $f_\delta = \mathrm{sign}(f)(|f| \wedge k_\delta)$. Because $k_\delta$ is almost everywhere finite, we have almost everywhere convergence in mean of $X_\delta^{(n)}$ to $f_\delta$. Now we can apply the Lebesgue dominated convergence theorem to the sequence of functions $\mathrm{E}|X_\delta^{(n)} - f_\delta|$ (bounded by $2k_\delta$) and conclude that $\int_0^\tau \mathrm{E}|X_\delta^{(n)} - f_\delta|$ converges to zero. From this, we may conclude that $\int_0^\tau X_\delta^{(n)}$ converges in mean, and hence in probability, to $\int_0^\tau f_\delta$. Finally, we let $\delta$ converge to zero to get the required result, for the probability that $\int_0^\tau X_\delta^{(n)} \neq \int_0^\tau X^{(n)}$ is less than or equal to $\delta$, whereas $\int_0^\tau f_\delta \to \int_0^\tau f$ as $\delta \to 0$ by monotone convergence.

## II.5.2. Further Results

Here we collect some further results often useful in proving limit theorems, though not all of them specifically asymptotic in nature. The first of them, the inequality of Lenglart (1977), will be used time and time again, especially in verifying the conditions of the martingale central limit theorem. In fact, it serves as an important lemma in the very proof of that theorem (Rebolledo, 1980a). The other results, concerning stochastic time change and Kurtz' (1983) results on convergence of a suitably rescaled counting process to a deterministic function, will only be used for some special cases: the pornoscope example (Examples I.3.16 and IV.1.10), the software reliability example (Examples I.3.19 and VI.1.12), and the Total Time on Test Plot (Section VI.3.2). The pornoscope and software reliability examples have in common that the counting processes in these models are defined directly by specifying their intensity processes rather than being built up by aggregation of simple independent (and often identically distributed) components. In the latter case, the Glivenko–Cantelli theorem and similar empirical process theory will provide the "convergence in probability" assumptions of the martingale central limit theorem. The Total Time on Test Plot is inspired directly by stochastic time change theory.

### II.5.2.1. *Lenglart's Inequality*

Suppose $\tilde{X}$ is the (nondecreasing) compensator of a local submartingale $X$. It turns out that we can bound the probability of a large value of $X$ anywhere in the whole time interval $\mathscr{T}$ in terms just of the probability of a large value of $\tilde{X}$ in the endpoint $\tau$. One says that $\tilde{X}$ dominates $X$. The inequality of Lenglart (1977) is then, for any $\eta > 0$ and $\delta > 0$,

$$\mathrm{P}\left(\sup_{\mathscr{T}} X > \eta\right) \leq \frac{\delta}{\eta} + \mathrm{P}(\tilde{X}(\tau) > \delta). \tag{2.5.17}$$

Special cases are that $X$ is $\int H \, \mathrm{d}N$, the integral of a non-negative predictable process $H$ with respect to a counting process $N$ (dominated by $\int H \, \mathrm{d}\Lambda$), and that $X$ is the square of a local square integrable martingale $M$ (dominated by $\langle M \rangle$). In the latter case, the inequality can be rewritten as

$$\mathrm{P}\left(\sup_{\mathscr{T}} |M| > \eta\right) \leq \frac{\delta}{\eta^2} + \mathrm{P}(\langle M \rangle(\tau) > \delta). \tag{2.5.18}$$

Thus, if $\langle M \rangle(\tau)$ is small, $M$ is small in absolute value throughout $\mathscr{T}$.

### II.5.2.2. *Stochastic Time Change*

Suppose $\mathbf{N}$ is a multivariate counting process with continuous compensator $\mathbf{\Lambda}$ satisfying $\Lambda_h(\tau) = \infty$ almost surely for each $h$. Let $\mathbf{\Lambda}^{-1}$ be the vector of

right-continuous inverses $\Lambda_h^{-1}$, defined on the new time interval $[0, \infty)$. We define the time-transformed multivariate counting process

$$\tilde{\mathbf{N}} = \mathbf{N} \circ \mathbf{\Lambda}^{-1} \tag{2.5.19}$$

as the vector of individually transformed processes $N_h \circ \Lambda_h^{-1}$. Graphically, this time transformation corresponds to plotting $N_h$ against its compensator $\Lambda_h$ instead of against the natural time coordinate. The surprising result is that $\tilde{\mathbf{N}}$ is equal in distribution to a vector $\mathbf{\Pi}$ of *independent* unit rate Poisson processes (see Examples II.3.1 and II.3.2).

If the condition $\Lambda_h(\tau) = \infty$ almost surely does not hold, one can still get a partial result by appending an independent unit rate Poisson process to $\tilde{N}_h$ from the time $\Lambda_h(\tau)$. In this way, one can see that $\tilde{N}_h$ is equal to $\Pi_h$, stopped at the random time $\Lambda_h(\tau)$. In the univariate case, $\Lambda(\tau)$ is a stopping time relative to the time-transformed filtration $(\mathcal{F}_{\Lambda^{-1}(u)})_{u \geq 0}$, but, in general, no such simple property holds.

These results are related to results on general random time transformations of a counting process. Consider a multivariate counting process $\mathbf{N}$ on the closed time interval $[0, \tau]$ with intensity process $\boldsymbol{\lambda}$, and let $R$ be a nondecreasing, adapted, continuous process. It is, therefore, predictable. The times $R^{-1}(r) = \inf\{t: R(t) \geq r\}$ are stopping times, and by applying Doob's optional stopping theorem to (a localized version of) the counting process local martingale $\mathbf{M} = \mathbf{N} - \int \boldsymbol{\lambda}$ and these stopping times, one finds that $\mathbf{M} \circ R^{-1}$ is a vector of local martingales with respect to the time-transformed filtration $(\mathcal{F}_{R^{-1}(r)})_{r \geq 0}$. Because $\mathbf{N} \circ R^{-1}$ is a multivariate counting process with respect to this filtration, uniqueness of compensators shows that its compensator is $\int_0^{R^{-1}(\cdot)} \boldsymbol{\lambda}$. If the paths of $R$ are absolutely continous with density $R'$, the intensity process of $\mathbf{N} \circ R^{-1}$ is, therefore, $(\boldsymbol{\lambda}/R') \circ R^{-1}$. The transformed counting process and its intensity are "stopped" at the stopping time $R(\tau)$.

Random time transformation will be used in Section VI.3.2 on the total time on test plot. The results on transformation to a Poisson process lead to the following topic: Kurtz' (1983) limit theorems for self-exciting counting processes.

### II.5.2.3. *Kurtz' Theorems*

Consider a multivariate counting process $\mathbf{N}$ with its self-exciting filtration $(\mathcal{N}_t)$. The compensator of $\mathbf{N}$, being predictable with respect to the filtration generated by $\mathbf{N}$, must be a "nonanticipating" functional of $\mathbf{N}$; that is, for each $t$, $\mathbf{\Lambda}(t)$ is a certain function (depending on $t$) of the path of $\mathbf{N}$ up to but not including time $t$. Write $\mathbf{N}^{t-}$ for the process $\mathbf{N}$ stopped "just before time $t$," i.e., $\mathbf{N}^{t-}(s) = \mathbf{N}(s)$ for $s < t$ and $\mathbf{N}^{t-}(s) = \mathbf{N}(t-)$ for $s \geq t$. With this notation, the nonanticipating nature of $\mathbf{\Lambda}$ means that we can consider $\mathbf{\Lambda}(t)$, not as a function of $t$ and of $\omega \in \Omega$, but as a function of $t$ and of $\mathbf{N}^{t-}(\cdot, \omega)$.

Random time transformation to a Poisson process now implies that in the

self-exciting case, one can actually construct $\mathbf{N}$ in distribution by starting with a vector of independent unit rate Poisson processes $\mathbf{\Pi}$ and then defining $\mathbf{N}$ pathwise as the solution of the implicit equation

$$\mathbf{N} \circ \mathbf{\Lambda}^{-1} = \mathbf{\Pi}, \tag{2.5.20}$$

cf. (2.5.19). Intuitively speaking, given a realisation of $\mathbf{\Pi}$, if we have already constructed the path of $\mathbf{N}$ up to time $t-$, then, in the case when $\mathbf{N}$ has an intensity process $\boldsymbol{\lambda}$, we can calculate $\Lambda_h(t)$ and $\lambda_h(t)$ for each $h$ from the path of $\mathbf{N}$ so far, and then define $N_h(t + \mathrm{d}t) = \Pi_h(\Lambda_h(t) + \lambda_h(t)\,\mathrm{d}t)$. Kurtz (1983) gave conditions on $\mathbf{\Lambda}(\cdot;\mathbf{N})$ which guarantee that the solution of (2.5.20) is unique.

Furthermore, limit theorems can be obtained using the fact that for a Poisson process, the law of large numbers and the central limit theorem hold. Let $\boldsymbol{\iota}$ be the vector of identity mappings $\iota_h(t) = t$ and let $\mathbf{\Pi}^{(n)}$ be the process $\mathbf{\Pi}^{(n)}(t) = \mathbf{\Pi}(nt)$. Then we have that, almost surely, $\mathbf{\Pi}^{(n)}/n$ converges uniformly on bounded intervals to $\boldsymbol{\iota}$ (the strong law of large numbers) and that $\sqrt{n}(\mathbf{\Pi}^{(n)}/n - \boldsymbol{\iota})$ converges in distribution to a vector of independent standard Wiener processes (see Example II.3.3).

These results are applied by defining a sequence of counting processes $\mathbf{N}^{(n)}$ by the modification of (2.5.20):

$$\mathbf{N}^{(n)} \circ \mathbf{\Lambda}^{-1}(\cdot;\mathbf{N}^{(n)}) = \mathbf{\Pi}^{(n)}. \tag{2.5.21}$$

It turns out that under suitable smoothness conditions on the functional dependence of $\mathbf{\Lambda}$ on $\mathbf{N}$, both law of large numbers and central limit theorem carry over from $\mathbf{\Pi}^{(n)}$ to $\mathbf{N}^{(n)}$. In the following version of the law of large numbers (all that we will have use for later), we describe a sequence of counting processes $\mathbf{N}^{(n)}$ corresponding to (2.5.21), as well as the limiting function to which $\mathbf{N}^{(n)}/n$ converges.

**Theorem II.5.4** [Kurtz' (1983) law of large numbers]. *Let* $\beta_h(t, \mathbf{x}) = \beta_h(t, \mathbf{x}^{t-})$ *be nonanticipating non-negative functions of an element* $\mathbf{x} \in (D(\mathcal{T}))^k$ *and* $t \in \mathcal{T}$ *such that* $\sup_{s \le t} \beta_h(s, \mathbf{x}) \le C_1 + C_2 \sup_{s<t} |\mathbf{x}(s)|$ *and* $\sup_{s \le t} |\beta_h(s, \mathbf{x}) - \beta_h(s, \mathbf{y})| \le C \sup_{s<t} |\mathbf{x}(s) - \mathbf{y}(s)|$, *for all* $\mathbf{x}$, $\mathbf{y} \in (D(\mathcal{T}))^k$, *and for certain constants* $C_1$, $C_2$, *and* $C$. *Let* $a_n \to \infty$ *be a sequence of positive constants. Let* $\mathbf{N}^{(n)}$ *be the multivariate counting process with intensity process* $\boldsymbol{\lambda}^{(n)} = a_n \boldsymbol{\beta}(\cdot, a_n^{-1}\mathbf{N}^{(n)})$ *and let* $\mathbf{X}$ *be the unique solution of the equation* $\mathbf{X}(t) = \int_0^t \boldsymbol{\beta}(s, \mathbf{X})\,\mathrm{d}s$ *for all* $t$. *Then* $\sup_{s \le t} |a_n^{-1}\mathbf{N}^{(n)} - \mathbf{X}| \xrightarrow{\mathrm{P}} 0$ *as* $n \to \infty$ *for all* $t \in \mathcal{T}$.

## II.6. Product-Integration and Markov Processes

In Section II.1, we saw the informal use of the *product-integral* both in formulas for likelihood functions and in the representation of a survival functional in terms of its cumulative hazard function. The notation used here, $\pi$,

is supposed to suggest a continuous version of the ordinary product $\prod$, just as the integral $\int$ generalizes the sum $\sum$ (see Gill and Johansen, 1990).

In the next section, we will take up the first topic and use product-integration to represent the likelihood function of a process observed on a time interval $[0, \tau]$ as an infinite product of conditional likelihoods for the development of the process in each infinitesimal time interval $[t, t + \mathrm{d}t)$, $t \in [0, \tau]$, given its past history. Exactly the same construction appears in the relationship between a *survival function* and its *hazard rate* or, more generally, its *hazard measure*. In multistate event history analysis, this relationship becomes that between the matrix of infinitesimal intensities or *transition rates* and the *transition probability matrix* of a continuous-time *Markov process*. These relationships will be reviewed below. They are central to our study of the Kaplan–Meier estimator and its generalization to Markov processes in Sections IV.3 and IV.4.

Though statistical models are often specified for continuous-time processes, natural statistical estimators are often discrete in nature. This makes it important to be able to handle discrete and continuous versions of the above relationship within a uniform framework. In fact, the basic and simple analytic tool of product-integration makes the discrete-time multiplication of a series of Markov one-step transition matrices and the continuous-time solution of the Kolmogorov differential equations two special cases of the product-integration of the matrix *intensity measure* of a Markov process.

The following results are taken from Gill and Johansen (1990), where complete proofs are given. Further background comments are given in the bibliographic comments to this chapter. In Section X.3.1, we will look at product-integration from a different point of view, leading to extensions to multivariate time.

**Definition II.6.1.** Let $\mathbf{X}(t)$, $t \in \mathcal{T}$, be a $p \times p$ matrix of cadlag functions of locally bounded variation. We define

$$\mathbf{Y} = \prod (\mathbf{I} + \mathrm{d}\mathbf{X}),$$

the product-integral of $\mathbf{X}$ over intervals of the form $[0, t]$, $t \in \mathcal{T}$, as the following $p \times p$ matrix function:

$$\mathbf{Y}(t) = \prod_{s \in [0,t]} (\mathbf{I} + \mathbf{X}(\mathrm{d}s)) = \lim_{\max |t_i - t_{i-1}| \to 0} \prod (\mathbf{I} + \mathbf{X}(t_i) - \mathbf{X}(t_{i-1})), \quad (2.6.1)$$

where $0 = t_0 < t_1 < \cdots < t_n = t$ is a partition of $[0, t]$ and the matrix product is taken in its natural order from left to right. In the leftmost term of the product, $\mathbf{X}(0)$ must be replaced by $\mathbf{X}(0-) = \mathbf{0}$ because the left endpoint 0 is included in the interval $[0, t]$.

We similarly define the product-integral over an arbitrary subinterval of $\mathcal{T}$, taking care of the endpoints in the natural way. The first result on product-integration is that the limit, indeed, always exists.

When the function $\mathbf{X}$ is a step-function, i.e., its components are the distribution functions of discrete measures, the product-integral becomes just a finite product over the jump times of $\mathbf{X}$ of the identity matrix plus the jumps of $\mathbf{X}$; thus,

$$\mathbf{Y} = \prod (\mathbf{I} + \Delta\mathbf{X}).$$

In the scalar case, $p = 1$, the order of multiplication does not matter, and one can separate the jumps of $X$ from its continuous part and get

$$\pi (1 + dX) = \exp(X^c) \prod (1 + \Delta X), \tag{2.6.2}$$

where $X^c = X - \sum \Delta X$ and $\Delta X = X - X_-$. So if $X$ is not only scalar but also continuous (and of bounded variation), the product-integral is just the ordinary exponential $\pi (1 + dX) = \exp(X)$.

A most important property of product-integration is the *multiplicativity* of product-integrals over disjoint intervals, which follows easily from the definition: for $0 \leq s \leq t \leq u$, we have

$$\underset{(s,u]}{\pi} (\mathbf{I} + d\mathbf{X}) = \underset{(s,t]}{\pi} (\mathbf{I} + d\mathbf{X}) \underset{(t,u]}{\pi} (\mathbf{I} + d\mathbf{X}).$$

Not only does the product-integral exist, but it is also the unique solution of a certain integral equation. This is why it was introduced by Volterra (1887).

**Theorem II.6.1.** $\pi (\mathbf{I} + d\mathbf{X})$ *exists and is (componentwise) a cadlag function of locally bounded variation. It is the unique solution to the integral equation*

$$\mathbf{Y}(t) = \mathbf{I} + \int_{s \in [0,t]} \mathbf{Y}(s-)\mathbf{X}(ds). \tag{2.6.3}$$

This result, like most of the results on product-integration, is actually rather intuitive. The right-hand side of (2.6.1) is the solution of the naturally associated discrete difference scheme for numerically solving (2.6.3) [more formally, it is the result of applying the first-order Euler scheme to the numerical solution of (2.6.3)]. We will see (2.6.3) shortly also as an integral version of the *Kolmogorov forward equation* for a Markov process.

Next we give a few key results on product-integration. The first one, Duhamel's equation, can be thought of as the continuous version of the elementary equality

$$\prod_{i=1}^{n} (\mathbf{I} + \mathbf{A}_i) - \prod_{i=1}^{n} (\mathbf{I} + \mathbf{B}_i) = \sum_{i=1}^{n} \left( \prod_{j=1}^{i-1} (\mathbf{I} + \mathbf{A}_j)(\mathbf{A}_i - \mathbf{B}_i) \prod_{j=i+1}^{n} (\mathbf{I} + \mathbf{B}_j) \right),$$

valid for matrices $\mathbf{A}_i$, $\mathbf{B}_i$, $i = 1, \ldots, n$.

**Theorem II.6.2.** *Let* $\mathbf{Y} = \pi (\mathbf{I} + d\mathbf{X})$, $\mathbf{Y}' = \pi (\mathbf{I} + d\mathbf{X}')$. *Then*

$$\mathbf{Y}(t) - \mathbf{Y}'(t) = \int_{s \in [0,t]} \underset{[0,s)}{\pi} (\mathbf{I} + d\mathbf{X})(\mathbf{X}(ds) - \mathbf{X}'(ds)) \underset{(s,t]}{\pi} (\mathbf{I} + d\mathbf{X}'). \tag{2.6.4}$$

When $\mathbf{Y}'(t)$ is nonsingular, one may right-multiply throughout in (2.6.4) by $\mathbf{Y}'(t)^{-1} = (\prod_{(s,t]}(\mathbf{I} + d\mathbf{X}'))^{-1}\mathbf{Y}'(s)^{-1}$ by multiplicativity. This leads to the version of (2.6.4) we will use most often:

$$\mathbf{Y}(t)\mathbf{Y}'(t)^{-1} - \mathbf{I} = \int_{s\in[0,t]} \prod_{[0,s)} (\mathbf{I} + d\mathbf{X})(\mathbf{X}(ds) - \mathbf{X}'(ds))\left(\prod_{[0,s]} (\mathbf{I} + d\mathbf{X}')\right)^{-1}$$

$$= \int_0^t \mathbf{Y}(s-)(\mathbf{X}(ds) - \mathbf{X}'(ds))\mathbf{Y}'(s)^{-1}. \tag{2.6.5}$$

In Section II.8, we give functional *continuity* and even *differentiability* results on product-integration derived from the Duhamel equation.

The next result generalizes Theorem II.6.1, showing that the product-integral arises in the solution of a whole class of integral equations.

**Theorem II.6.3.** *Let* $\mathbf{Z}$, $\mathbf{W}$ *be* $k \times p$ *matrix cadlag functions. For given* $\mathbf{W}$, *the unique solution* $\mathbf{Z}$ *of the Volterra equation*

$$\mathbf{Z}(t) = \mathbf{W}(t) + \int_0^t \mathbf{Z}(s-)\mathbf{X}(ds) \tag{2.6.6}$$

*is*

$$\mathbf{Z}(t) = \mathbf{W}(t) + \int_0^t \mathbf{W}(s-)\mathbf{X}(ds) \prod_{(s,t]} (\mathbf{I} + d\mathbf{X})$$

$$= \mathbf{W}(0) \prod_{[0,t]} (\mathbf{I} + d\mathbf{X}) + \int_0^t \mathbf{W}(ds) \prod_{(s,t]} (\mathbf{I} + d\mathbf{X}). \tag{2.6.7}$$

Finally, we give two results which are sometimes useful, one giving an explicit series representation of the product-integral, and the other on its determinant in the continuous case:

**Theorem II.6.4.** [Péano Series, Péano (1888)].

$$\prod_{[0,t]} (\mathbf{I} + d\mathbf{X}) = \mathbf{I} + \sum_{n=1}^{\infty} \int\cdots\int_{0\le s_1<\cdots<s_n\le t} \mathbf{X}(ds_1)\cdots\mathbf{X}(ds_n). \tag{2.6.8}$$

**Theorem II.6.5.** *Suppose* $d\mathbf{X}(t) = \mathbf{U}(t)\,dt$ *and* $\mathbf{X}(0) = \mathbf{0}$. *Then*

$$\det \prod_{[0,t]} (\mathbf{I} + d\mathbf{X}) = \exp\left(\int_0^t \operatorname{trace} \mathbf{U}(s)\,ds\right). \tag{2.6.9}$$

Next we show how product-integration arises in the relation between survival functions and hazard measures.

**Theorem II.6.6.** *Let* $S$ *be a survival function of a positive random variable (survival time)* $T$, *i.e.,* $S(t) = \mathrm{P}(T > t)$ *for all* $t \ge 0$ *and* $S(0) = 0$. *Define the cumulative or integrated hazard function*

$$A(t) = -\int_0^t \frac{S(\mathrm{d}s)}{S(s-)}. \tag{2.6.10}$$

*Then*

$$S(t) = \prod_{[0,t]} (1 - \mathrm{d}A), \tag{2.6.11}$$

*for all t such that* $A(t) < \infty$.

PROOF. Note that for all $t$ such that $S(t-) > 0$ and, hence, $A(t) < \infty$,

$$S(t) = 1 - \int_0^t S(s-)A(\mathrm{d}s).$$

Now apply Theorem II.6.1 with $X = -A$ and $\mathscr{T} = \{t: S(t-) > 0\}$. If $\mathscr{T} = [0, \tau]$ for some $\tau < \infty$, then $S(\tau-) > 0$ but $S(\tau) = 0$. Consequently, $A(t) = A(\tau) < \infty$ for all $t \geq \tau$ and the relation (2.6.11) holds for all $t \geq \tau$ too. If, however, $\mathscr{T} = [0, \tau)$ for some $\tau \leq \infty$, one can show by some straightforward analysis that $A(t) = A(\tau) = \infty$ for all $t \geq \tau$ and (2.6.11) cannot be extended (except by way of definition) to $t \geq \tau$. □

When $S$ is absolutely continuous so that $F = 1 - S$ has density $f$, we define the hazard rate $\alpha = f/(1 - F)$. The integrated hazard is then $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ and relation (2.6.11) by (2.6.2) becomes

$$S(t) = \prod_0^t (1 - \alpha(s)\,\mathrm{d}s) = \exp\left(-\int_0^t \alpha(s)\,\mathrm{d}s\right). \tag{2.6.12}$$

If, however, $S$ is discrete, one can define the discrete hazard function $\alpha(t) = \mathrm{P}(T = t \mid T \geq t)$. Then $A(t) = \sum_{s \leq t} \alpha(s)$ and (2.6.11) becomes

$$S(t) = \prod_{s \leq t} (1 - \alpha(s)).$$

The cumulative hazard function arose in the compensator of the counting process registering, by a jump at time $T$, the end of the survival time, cf. Example II.4.1 and the remarks following this example. Let $T$ be a survival time and define $\mathscr{F}_t = \sigma\{T \wedge t, I(T \leq t)\}$; let $N(t) = I(T \leq t)$ and $Y(t) = I(T \geq t)$. Then $N$ has compensator $\Lambda$ given by

$$\Lambda(t) = \int_0^t Y(s)A(\mathrm{d}s).$$

Next, we extend these results to an almost arbitrary finite state-space, continuous-time Markov process. We explain the implied restriction afterward. First we need a definition. We say that a $p \times p$ matrix function $\mathbf{A}$ corresponds to a locally finite *intensity measure of a Markov process* on a time interval $\mathscr{T}$ if the $A_{hj}$, $h \neq j$, are nondecreasing cadlag functions, zero at time zero, $A_{hh} = -\sum_{j \neq h} A_{hj}$, and $\Delta A_{hh}(t) \geq -1$ for all $t$. The function $A_{hj}$ is called the integrated intensity function for transitions from state $h$ to state $j$,

whereas $A_{hh}$ is the negative integrated intensity function for transitions out of state $h$.

**Theorem II.6.7.** *Let the matrix function* $\mathbf{A}$ *correspond to an intensity measure. Define*

$$\mathbf{P}(s,t) = \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}), \quad s \le t;\ s, t \in \mathscr{T}. \tag{2.6.13}$$

*Then* $\mathbf{P}$ *is the transition matrix of a Markov process with state space* $\{1,\dots,p\}$ *and intensity measure* $\mathbf{A}$. *The process can be constructed as follows (starting from any time instant in any state): Given the process is in state* $h$ *at time* $t_0$, *it remains in this state for a length of time with integrated hazard function*

$$-(A_{hh}(t) - A_{hh}(t_0)), \quad t_0 \le t \le \inf\{u \ge t_0 \colon \Delta A_{hh}(u) = -1\}.$$

*Given that it jumps out of state* $h$ *at time* $t$, *it jumps into state* $j \neq h$ *with probability* $(\mathrm{d}A_{hj}/(-\mathrm{d}A_{hh}))(t)$.

The theorem is not proved here; one has to show that the construction described in the theorem really does define a process (making only finitely many jumps in finite time intervals), that this process is Markov, and that its transition matrices are given by (2.6.13).

An important special case is when the $A_{hj}$ are absolutely continuous, $A_{hj} = \int \alpha_{hj}$ for certain *intensity functions* or *transition rates* or *transition intensities* $\alpha_{hj}$. The sojourn times in a given state $h$ are then continuously distributed with hazard rate function $-\alpha_{hh}$ (starting from the time of entry into state $h$). The conditional jump probabilities (given a jump out of state $h$ at time $t$) equal $\alpha_{hj}(t)/(-\alpha_{hh}(t))$. We let $\boldsymbol{\alpha}$ be the matrix of these transition intensities. The integral equation (2.6.3) for the product-integral of $\mathbf{A}$ can be rewritten as an integral form of the Kolmogorov forward differential equations for the transition matrix $\mathbf{P}$ with (time varying) intensities $\alpha_{hj}$: $\mathbf{P}(s,t)$ is the unique solution of the equations

$$\mathbf{P}(s,s) = \mathbf{I},$$

$$\frac{\partial}{\partial t}\mathbf{P}(s,t) = \mathbf{P}(s,t)\boldsymbol{\alpha}(t)$$

because (starting at time $s$ instead of 0) Eq. (2.6.3) becomes

$$\mathbf{P}(s,t) = \mathbf{I} + \int_{u \in (s,t]} \mathbf{P}(s,u)\boldsymbol{\alpha}(u)\,\mathrm{d}u.$$

Another special case is when the $A_{hj}$ are step-functions. Now the Markov process is a discrete-time Markov chain and the matrix $\mathbf{I} + \Delta\mathbf{A}(t)$ is the transition matrix for a jump at time $t$. The product-integral (2.6.13) describes the transition matrix for the time interval $(s,t]$ as the product of the transition matrices for each possible jump time between $s$ and $t$ (including $t$). The sojourn times have discrete distributions with discrete hazard function $-\Delta A_{hh}$ for leaving state $h$. Given a jump out of $h$ occurs at time $t$, the

new state is state $j$ with probability $\Delta A_{hj}(t)/(-\Delta A_{hh}(t))$. For a discrete-time Markov process, one can define $\alpha_{hj}(t) = \mathrm{P}(X(t) = j | X(t-) = h)$, $h \neq j$, and $\alpha_{hh}(t) = -\mathrm{P}(X(t) \neq h | X(t-) = h)$. So, $1 + \alpha_{hh}(t) = \mathrm{P}(X(t) = h | X(t-) = h)$. Equation (2.6.13) becomes the product of the transition matrices $\mathbf{I} + \boldsymbol{\alpha}(u) = \mathbf{I} + \Delta\mathbf{A}(u)$ for each time instant $u$ in $(s, t]$ at which a transition is possible.

In some of the examples of Section IV.4.1, we will need a form of the Kolmogorov forward equation for the situation with general integrated intensity functions. Using the definition (2.6.13), the integral equation (2.6.3) can be rewritten, for fixed $s$ and for $t \geq s$, as

$$\begin{aligned} \mathbf{P}(s, s) &= \mathbf{I}, \\ \mathbf{P}(s, \mathrm{d}t) &= \mathbf{P}(s, t-)\mathbf{A}(\mathrm{d}t). \end{aligned} \tag{2.6.14}$$

Finite state-space Markov processes whose transition matrices are *not* product-integrals of finite intensity measures are those such that for some state $h$ and times $u$ and $v$, given the process is in state $h$ at time $u$, its future sojourn time in this state has infinite cumulative hazard up to time $v$. If one artificially kept such a state occupied by letting a new individual start in that state every time the state was vacated, an infinite number of jumps out of the state would occur in finite time. The original process is certain to leave the state in finite time, but does not certainly leave at any specific time instant. It would be all right to include such states in the theory provided care is taken that this possibility does not lead to a positive probability of an infinite number of jumps in a finite time interval. In particular, if the destination state $j$ for an infinite intensity $A_{hj}$ is *absorbing*, thus all intensities for leaving $j$ are zero, this is not a problem because on leaving $h$ for $j$ there is no possibility of reentering $h$ again.

The intensity measure reappears in the compensator of the counting processes registering each type of jump of the process (Jacobsen, 1982, p. 120). This is worth summarizing in a last theorem:

**Theorem II.6.8.** *Let* $\mathbf{A}$ *correspond to the intensity measure of a Markov process* $X$. *Let* $\mathcal{F}_t = \sigma\{X(s): s \leq t\}$; *define*

$$\begin{aligned} Y_h(t) &= I(X(t-) = h), \\ N_{hj}(t) &= \#\{s \leq t: X(s-) = h, X(s) = j\}, \quad h \neq j. \end{aligned}$$

*Then* $\mathbf{N} = (N_{hj}, h \neq j)$ *is a multivariate counting process and its compensator with respect to* $(\mathcal{F}_t) = (\sigma(X(0)) \vee \mathcal{N}_t)$ *has components*

$$\Lambda_{hj}(t) = \int_0^t Y_h(s) A_{hj}(\mathrm{d}s). \tag{2.6.15}$$

*Equivalently, the processes* $M_{hj}$ *defined by*

$$M_{hj} = N_{hj} - \int Y_h \, \mathrm{d}A_{hj}$$

*are martingales.*

In the absolutely continuous case with transition intensities $\alpha$, and $\mathbf{A}(t) = \int_0^t \alpha(s)\,ds$, (2.6.15) gives us that $\mathbf{N}$ has intensity $\lambda = (\lambda_{hj}: h \neq j)$ with

$$\lambda_{hj}(t) = Y_h(t)\alpha_{hj}(t).$$

The result on the intensity of the one-jump counting process following Theorem II.6.6 can be seen as a special case of Theorem II.6.8 by letting $T$ be the time of the only transition of a two-state Markov process from state 0 to state 1, the process starting in state 0 at time 0.

# II.7. Likelihoods and Partial Likelihoods for Counting Processes

## II.7.1. Jacod's Formulas

In this section, we will consider a multivariate counting process $\mathbf{N} = (N_1, \ldots, N_k)$ under various probability measures and, hence, with a varying (cumulative) intensity process. The counting process will be defined on a fixed probability space and filtration satisfying the usual conditions except possibly completeness (see Section II.2). We will work throughout under the special assumptions

$$\begin{gathered} \mathscr{F}_t = \mathscr{F}_0 \vee \sigma\{\mathbf{N}(s): s \leq t\} = \mathscr{F}_0 \vee \mathscr{N}_t, \\ \mathscr{T} = \overline{\mathscr{T}} = [0, \tau], \qquad \mathscr{F} = \mathscr{F}_\tau. \end{gathered} \tag{2.7.1}$$

This means, in particular, that $\mathbf{N}(\tau)$ and $\Lambda(\tau)$ are both finite.

We can think of a probability measure P on $\mathscr{F}$ as being built up in the following way. First, specify P on $\mathscr{F}_0$. Then, specify the conditional distribution, given $\mathscr{F}_0$, of the time $T_1 \in (0, \tau]$ of the first jump of $\mathbf{N}$. Next, specify the conditional distribution of the *type* of the first jump $J_1 \in \{0, 1, \ldots, k\}$ given $\mathscr{F}_0$ and $T_1$, where $J_1 = 0$ is only allowed if $T_1 = \tau$ (and means that there is no first jump of $\mathbf{N}$). Go on with a conditional distribution of $T_2 \in (T_1, \tau]$ given $\mathscr{F}_0$, $T_1$, $J_1$; and so on. The construction could only fail to define a proper counting process model through an explosion: Infinitely many jumps occur before time $\tau$ with positive probability. We would like to be sure that this cannot happen. However, the probability of an explosion is a very complicated functional of all the conditional distributions we have just described. We will typically have to make do with simple sufficient conditions (i.e., conditions which are much stronger than necessary). Theorems II.7.4 and II.7.5 illustrate the problem: The first (the *Girsanov theorem*) makes the weakest possible assumption, but it is quite impossible to verify without making strong boundedness conditions. The second of the theorems makes a much stronger assumption but, at least, it is easy to check.

Before coming to construction theorems, we consider alternative ways to

*describe* the distribution of a counting process. The conditional distributions of the $T_n$, given the past, can also be specified through their cumulative hazard functions. Multiplying the hazard rate at time $t$ by the probability of a jump of type $h$ at time $t$, given one occurs then, produces type-specific hazards; conversely, from $k$ type-specific hazards for $(T_n, J_n)$ given the past, we can reconstruct the total hazard and the conditional, time-dependent probability of an event of each type, given one occurs.

Now the type-specific hazards are, intuitively speaking, just the intensity processes of the components of $\mathbf{N}$. So this informal argument (see Section II.1 and also the discussion following Example II.4.1) suggests that under (2.7.1), the (cumulative) intensity process of $\mathbf{N}$ is determined by the conditional probability distributions just described and, conversely, from the intensity processes, we can recover the conditional distributions and, hence, together with knowledge of P on $\mathscr{F}_0$, reconstruct P on $\mathscr{F}$.

Instead of going from jump to jump (taking big steps through $\mathscr{T}$), one could also go from infinitesimal time interval to infinitesimal time interval, with the probability of a jump of type $h$ in the interval $[t, t + \mathrm{d}t)$, given the past ($\mathscr{F}_0$ and the path of $\mathbf{N}$ up to time $t$), being just $\mathrm{d}\Lambda_h(t)$. This gives an alternative way of writing down probability densities or likelihood ratios which may be easier to interpret and leads to the notion of *partial likelihood* in a natural way.

We now make these ideas rigorous by stating some theorems due to Jacod and others. Recall (cf. Section II.4.4) that we write ${}_TX$ for a process $X$ *started* at time $T$, i.e., ${}_TX(t) = X(t) - X^T(t)$.

**Theorem II.7.1** (Jacod's Formula for the Intensity Process). *Under* (2.7.1), *let* $P_n$ *be a regular version of the joint conditional distribution of* $(T_n, J_n)$ *given* $\mathscr{F}_{T_{n-1}} = \mathscr{F}_0 \vee \sigma\{T_1, J_1, \ldots, T_{n-1}, J_{n-1}\}$. *Then, on the time interval* $(T_{n-1}, T_n]$, *we have*

$$ {}_{T_{n-1}}\Lambda_h(t) = \int_{T_{n-1}}^{t} \frac{P_n(\mathrm{d}s, \{h\})}{P_n([s, \tau], \bar{E})}, \quad h = 1, \ldots, k, $$

*where* $E = \{1, \ldots, k\}$ *and* $\bar{E} = E \cup \{0\}$. *Conversely, on* $(T_{n-1}, T_n)$, *we have*

$$ P_n((t, \tau], \bar{E}) = \prod_{T_{n-1}}^{t} (1 - \mathrm{d}\Lambda_\cdot(s)) $$

($_\cdot$ *denoting summation over* $1, \ldots, k$) *and, with* $P_n(h|t)$, *the* $P_n$*-conditional distribution of* $J_n$ *given* $T_n = t$,

$$ P_n(h|t) = \frac{\mathrm{d}\Lambda_h(t)}{\mathrm{d}\Lambda_\cdot(t)}, \quad h = 1, \ldots, k. $$

This result is a generalization of Example II.4.1, the "one-jump process." The theorem also holds, with notational modification only, for a marked point process (introduced at the end of Section II.4.1): Replace the discrete "mark"

$h$ by an element $dx$ of a general mark space $(E, \mathscr{E})$ and interpret ratios of differentials as transition measures. We return to this in Section II.7.3.

To compute likelihood ratios, one could use the "converse" part of the preceding theorem to extract the joint conditional distribution of $T_1, J_1, T_2, J_2, \ldots$ given $\mathscr{F}_0$ from $\Lambda$. It is more convenient, however, to recast this in terms of the infinitesimal experiments described earlier. We use product-integral notation (Section II.6) together with some even more nonstandard notational conventions which are explained after the statement of the theorem.

**Theorem II.7.2** (Jacod's Formula for the Likelihood Ratio). *Suppose* (2.7.1) *holds and* $\mathrm{P}$ *and* $\tilde{\mathrm{P}}$ *are two probability measures on the filtered probability space under which* $\mathbf{N}$ *has compensators* $\Lambda$ *and* $\tilde{\Lambda}$, *respectively. Suppose* $\tilde{\mathrm{P}}$ *is absolutely continuous with respect to* $\mathrm{P}$, *written* $\tilde{\mathrm{P}} \ll \mathrm{P}$. *Then,*

$$\tilde{\Lambda}_h \ll \Lambda_h \quad \textit{for all } h, \ \mathrm{P}\textit{-a.s.},$$

$$\Delta\Lambda_\cdot(t) = 1 \quad \textit{for any } t \textit{ implies } \Delta\tilde{\Lambda}_\cdot(t) = 1, \ \mathrm{P}\textit{-a.s.}$$

*and*

$$\begin{aligned}\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}} &= \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}\bigg|_{\mathscr{F}_0} \frac{\pi_{t\in[0,\tau]}\left(\prod_h d\tilde{\Lambda}_h(t)^{\Delta N_h(t)}(1-d\tilde{\Lambda}_\cdot(t))^{1-\Delta N_\cdot(t)}\right)}{\pi_{t\in[0,\tau]}\left(\prod_h d\Lambda_h(t)^{\Delta N_h(t)}(1-d\Lambda_\cdot(t))^{1-\Delta N_\cdot(t)}\right)} \\ &= \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}\bigg|_{\mathscr{F}_0} \prod_t \prod_h \left(\frac{d\tilde{\Lambda}_h}{d\Lambda_h}(t)\right)^{\Delta N_h(t)} \frac{\pi_{t\in[0,\tau]:\Delta N_\cdot(t)\neq 1}(1-d\tilde{\Lambda}_\cdot(t))}{\pi_{t\in[0,\tau]:\Delta N_\cdot(t)\neq 1}(1-d\Lambda_\cdot(t))}. \end{aligned} \tag{2.7.2}$$

The first line of (2.7.2) is more intuitively easy to interpret: It describes the likelihood ratio as a ratio of products of conditional (and infinitesimal) multinomial experiments. The second line gives the intended formal mathematical meaning. The following remarks explain how to make the step between the two versions and can be considered as an algorithm for making sense of expressions like (2.7.2) in the future.

Because everything is real-valued, the order of the terms in the first version of (2.7.2) can be changed at will. The "jump parts" [i.e., those with superscript $\Delta N_h(t)$] only occur at a finite number of time points and can be taken as a discrete product outside the product-integrals. The product-integrals which are left are effectively taken over the subset of $t \in [0, \tau]$ for which $\Delta N_\cdot(t) = 0$, i.e., such that $t \neq T_n$ for any $n$. Because $\Delta\Lambda_\cdot(t) = 1 \Rightarrow \Delta N_\cdot(t) = 1$, P-a.s., this means we are not embarrassed by a term $(1 - \Delta\Lambda_\cdot(t))$ with $\Delta\Lambda_\cdot(t) = 1$ in the denominator. By our assumption (2.7.1), $\Lambda_\cdot(\tau) < \infty$, P-a.s., so the product-integral in the denominator is not zero. The product-integrals can finally, by (2.6.2), be evaluated as the ordinary products of the negative exponentials of the continuous parts of $\tilde{\Lambda}_\cdot$ and $\Lambda_\cdot$, together with the contributions $1 - \Delta\tilde{\Lambda}_\cdot$ and $1 - \Delta\Lambda_\cdot$ over the times in $[0, \tau]$ where $\mathbf{N}$ does not jump.

The jump parts in the first version of (2.7.2) should be intepreted by taking ratios and forming Radon–Nikodym derivatives $((d\tilde{\Lambda}_h/d\Lambda_h)(t))^{\Delta N_h(t)}$; according to the first statement of the theorem, this derivative is finite and we will have a finite number of such terms with $t = T_n$ and $h = J_n$ only.

Theorem II.7.2 also holds for marked point processes and we work with such a version in Section II.7.3.

Usually, we will have continuous or even absolutely continuous compensators. In these cases the product-integrals by (2.6.2) simplify to exponentials; and in the case with intensities, the Radon–Nikodym derivative of one *integrated intensity process* with respect to another becomes simply the ratio of the intensities. These cases are covered in the next corollary.

**Corollary II.7.3** (Continuous and Absolutely Continuous Case). *If* $\Lambda$ *and* $\tilde{\Lambda}$ *are* P*-a.s. continuous, then*

$$\frac{d\tilde{P}}{dP} = \frac{d\tilde{P}}{dP}\bigg|_{\mathcal{F}_0} \frac{\prod_{h,t} d\tilde{\Lambda}_h(t)^{\Delta N_h(t)} \exp(-\tilde{\Lambda}_\cdot(\tau))}{\prod_{h,t} d\Lambda_h(t)^{\Delta N_h(t)} \exp(-\Lambda_\cdot(\tau))} \tag{2.7.3}$$

*and if they are actually absolutely continuous, then*

$$\frac{d\tilde{P}}{dP} = \frac{d\tilde{P}}{dP}\bigg|_{\mathcal{F}_0} \frac{\prod_{h,t} \tilde{\lambda}_h(t)^{\Delta N_h(t)} \exp(-\tilde{\Lambda}_\cdot(\tau))}{\prod_{h,t} \lambda_h(t)^{\Delta N_h(t)} \exp(-\Lambda_\cdot(\tau))}. \tag{2.7.4}$$

The products in (2.7.4) are just $\prod_n \tilde{\lambda}_{J_n}(T_n)$ and $\prod_n \lambda_{J_n}(T_n)$. [A result of Brémaud (1981) states that a predictable intensity $\lambda_h$ is a.s. unique on the jump times of $N_h$. So $d\tilde{P}/dP$ *is* well defined by this expression.]

In statistical applications, we study *likelihood ratios* formed by taking Radon–Nikodym derivatives of members of the family of probability measures under consideration with respect to one fixed reference distribution. Likelihoods are only needed up to a proportionality factor. This means that one need only specify the numerators of the likelihood ratios in (2.7.2)–(2.7.4); we have indicated how the resulting formal expressions can be manipulated by formal algebra to give well-defined likelihood functions. So (dropping the tildes), we can rephrase the results of Theorem II.7.2 and Corollary II.7.3 as

$$dP = dP|_{\mathcal{F}_0} \prod_{t\in[0,\tau]} \left( \prod_h d\Lambda_h(t)^{\Delta N_h(t)} (1 - d\Lambda_\cdot(t))^{1-\Delta N_\cdot(t)} \right), \tag{2.7.2'}$$

$$\begin{aligned} dP &= dP|_{\mathcal{F}_0} \prod_{h,t} d\Lambda_h(t)^{\Delta N_h(t)} \prod_0^\tau (1 - d\Lambda_\cdot) \\ &= dP|_{\mathcal{F}_0} \prod_{h,t} d\Lambda_h(t)^{\Delta N_h(t)} \exp(-\Lambda_\cdot(\tau)), \end{aligned} \tag{2.7.3'}$$

$$\begin{aligned} dP &\propto dP|_{\mathcal{F}_0} \prod_{h,t} \lambda_h(t)^{\Delta N_h(t)} \prod_0^\tau (1 - d\Lambda_\cdot) \\ &= dP|_{\mathcal{F}_0} \prod_{h,t} \lambda_h(t)^{\Delta N_h(t)} \exp(-\Lambda_\cdot(\tau)), \end{aligned} \tag{2.7.4'}$$

for the general, continuous, and absolutely continuous case, respectively. In going from (2.7.3') to (2.7.4'), we have substituted $d\Lambda_h(t) = \lambda_h(t)\,dt$ and dropped the factors $dt$ which cancel anyway on forming ratios.

An even more informal way of writing the first of these equations will also be useful. Here, we simply replace $\Delta N_h(t)$, $\Delta N_\cdot(t)$ everywhere by $dN_h(t)$, $dN_\cdot(t)$:

$$dP = dP|_{\mathcal{F}_0} \prod_{t\in[0,\tau]} \left( \prod_h d\Lambda_h(t)^{dN_h(t)} (1 - d\Lambda_\cdot(t))^{1-dN_\cdot(t)} \right). \qquad (2.7.2'')$$

The intended mathematical interpretation is not changed; however, (2.7.2″) especially carries the suggestion that one might be able to calculate a Radon–Nikodym derivative $d\tilde{P}/dP$ by partitioning $[0, \tau]$ into small subintervals, computing increments of $\mathbf{N}$, $\mathbf{\Lambda}$, and $\tilde{\mathbf{\Lambda}}$ over the subintervals and forming the ratio of corresponding finite products; then one takes the limit as the partition becomes finer and finer.

Conditions can be written down under which the Radon–Nikodym derivative $d\tilde{\Lambda}_h/d\Lambda_h$ is indeed a limit of approximating discrete ratios, and the product-integral is actually *defined* as a limit of approximating finite products. However, a rigorous derivation (under appropriate conditions on the partitions) is not important for us because we are more concerned with the suggestive nature of the formula. This corresponds to the interpretation of $d\mathbf{\Lambda}(t)$ as the vector of conditional probabilities that $N_h$ has a jump in the time interval $[t, t + dt)$, $h = 1, \ldots, k$, given the past $\mathcal{F}_{t-}$. The likelihood is correspondingly written as a product of conditional multinomial probabilities for the infinitesimal subexperiments "in $[t, t + dt)$ observe $d\mathbf{N}(t)$."

We next make a technical comment on Theorem II.7.2 and Corollary II.7.3 concerning the condition that $\tilde{P}$ be absolutely continuous with respect to P. A sufficient condition for $\tilde{P} \ll P$ in the case when intensity processes exist is that $\tilde{\lambda}_h = 0$ where $\lambda_h = 0$, $\tilde{P}$-almost surely, and $\tilde{P}(\Lambda_\cdot(\tau) < \infty) = 1$. More complicated necessary and sufficient conditions are also available, due to Kabanov, Liptser, and Shiryayev (1976). However, because all these conditions are formulated in terms of conditions to hold $\tilde{P}$-almost surely, they are not of much help in *defining* $\tilde{P}$ from P via an expression for $d\tilde{P}/dP$.

In statistical applications, we will want to consider a whole family of probability measures P, not necessarily mutually absolutely continuous, and, therefore, cannot always apply Theorem II.7.2 to obtain $d\tilde{P}/dP$ for each $\tilde{P}$, P considered; this should take the value $+\infty$ on the $\tilde{P}$-singular part of $\Omega$ with respect to P. However, for any two probability measures $\tilde{P}$ and P, there exists a third, say, $Q = \frac{1}{2}(\tilde{P} + P)$, which dominates both P and $\tilde{P}$. We can, therefore, calculate $dP/dQ$ and $d\tilde{P}/dQ$ by Theorem II.7.2 and finally put

$$\frac{d\tilde{P}}{dP} = \frac{d\tilde{P}}{dQ} \Big/ \frac{dP}{dQ} \quad \text{where } \frac{dP}{dQ} > 0,$$

$$\frac{d\tilde{P}}{dP} = \infty \quad \text{where } \frac{dP}{dQ} = 0.$$

Note that $dP/dQ + d\tilde{P}/dQ = 2$, so we never have $dP/dQ = \infty$. Because the denominators of $d\tilde{P}/dQ$ and $dP/dQ$ given by (2.7.2) coincide, this comes down to calculating $d\tilde{P}/dP$ by (2.7.2) as it stands; the result is P-a.s. finite but possibly infinite with positive $\tilde{P}$ probability. Because

$$E_P\left(\frac{d\tilde{P}}{dP}\right) = E_P\left(\frac{d\tilde{P}}{dP} I\left(\frac{d\tilde{P}}{dP} < \infty\right)\right) = \tilde{P}(A),$$

where $A = \{d\tilde{P}/dP < \infty\}$, we have $\tilde{P} \ll P$ if and only if

$$E_P\left(\frac{d\tilde{P}}{dP}\right) = 1.$$

From this last result, we can now derive a result, called a Girsanov-type theorem by analogy with a similar technique for constructing diffusion processes, on the construction of a counting process with a given intensity process by reference to a given counting process model. (We do not use the theorem in this book but the Girsanov theorem is too famous to be omitted.) We state this in the case when intensities exist (in fact, the "reference" probability is most often in applications chosen to correspond to a standard Poisson process $\Pi$ on a finite time interval $[0, \tau]$):

**Theorem II.7.4** (Girsanov). *Suppose* $\mathbf{N}$ *has intensity process* $\lambda$ *under* P *and* (2.2.1) *holds. Let* $\tilde{\lambda}$ *be any predictable process whose components are* P*-a.s. zero where those of* $\lambda$ *are zero. Let* $\tilde{P}_0$ *be a given probability measure on* $\mathcal{F}_0$*, absolutely continuous with respect to* P *restricted to this sub-*$\sigma$*-algebra. Let* $Z$ *denote the right-hand side of* (2.7.4)*. Then there exists a probability measure* $\tilde{P}$ *on* $\mathcal{F}$*, which is absolutely continuous with respect to* P*, agrees with* $\tilde{P}_0$ *on* $\mathcal{F}_0$*, and is such that* $\mathbf{N}$ *has intensities* $\tilde{\lambda}$ *with respect to* $\tilde{P}$*, if and only if*

$$E_P(Z) = 1. \tag{2.7.5}$$

Note that we automatically have $E_P(Z) \leq 1$. Unfortunately, there is no easy general way to check (2.7.5). Brémaud (1981) gave some sufficient conditions which are sometimes useful.

Another construction result due to Jacobsen (1982) is possible in which a reference probability is not needed, provided we work with a "canonical" choice of $\Omega$. Also, we make some strong boundedness assumptions about the intensities. We will use this construction result for the pornoscope example (Examples I.3.16 and III.1.10) and the software reliability example (Examples I.3.19 and III.1.12), both of which are most conveniently specified through their intensity processes. First, we give this in the case when $\mathcal{F}_0$ is trivial. Let $\Omega$ be the collection of all possible sample paths of a multivariate counting process on $[0, \tau]$, i.e., the collection of all functions from $[0, \tau]$ to $\mathbb{N}_0^k$, zero at time zero, piecewise constant, right continuous, with jumps of size $+1$ only, no two components having simultaneous jumps. ($\mathbb{N}_0$ denotes the non-negative integers here.) The multivariate counting process $\mathbf{N}$ is defined on this $\Omega$ simply by $\mathbf{N}(\omega) = \omega$. On $\Omega$ we impose the filtration $(\mathcal{N}_t)$, $\mathcal{N}_t = \sigma\{\mathbf{N}(s): s \leq t\}$.

The idea of the construction is to specify the intensity process of the counting process as a function of its own past (past jump times and types) as well as the present time instant. So one must specify a large family of functions giving the intensity at each time $t$, for each number $n$ of previous

jumps, and for each precise set of jump times $t_i$ and types $j_i$, $i = 1, \ldots, n$. Assuming the integrated intensity is finite for any conceivable realization of $\mathbf{N}$ saves us from the possibility of explosion:

**Theorem II.7.5** (Jacobsen's Construction). *Suppose* $\lambda_n(t; t_1, \ldots, t_n; j_1, \ldots, j_n)$ *are given measurable non-negative k-variate functions defined for* $0 < t_1 < t_2 < \cdots < t_n \leq t \leq \tau, j_1, \ldots, j_n \in \{1, \ldots, k\}$. *For a given path* $\mathbf{N}$, *define*

$$\lambda(t; \mathbf{N}) = \lambda_n(t; t_1, \ldots, t_n; j_1, \ldots, j_n)$$

*if* $\mathbf{N}$ *has* $T_i$, $J_i$ *satisfying* $T_i = t_i$, $J_i = j_i$; $i = 1, \ldots, n$; $T_{n+1} \geq t$. *Suppose* $\int_0^\tau \lambda(t; \mathbf{N})\, dt < \infty$ *for all* $\mathbf{N}$. *Then we can define* $\mathbf{P}$ *on* $\Omega$ *such that* $\mathbf{N}$ *has intensity process* $\lambda(\mathbf{N}) = \lambda(t; \mathbf{N})$ *with respect to* $(\mathcal{N}_t)$.

By including dependence on an element $z$ of a measurable space $(\mathcal{Z}, \mathcal{B})$ in $\lambda_n$ and adding a probability distribution $P_0$ on $\mathcal{Z}$, one can extend this in the obvious way to a probability on the filtration $\mathcal{F}_t = \mathcal{B} \otimes \mathcal{N}_t$ on $\mathcal{Z} \times \Omega$.

## II.7.2. Martingale Properties of Likelihood Processes

Suppose (2.7.1) holds for each of $\tilde{\mathbf{P}}$ and $\mathbf{P}$, $\tilde{\mathbf{P}} \ll \mathbf{P}$, on a common filtration. Introduce the *likelihood process* $L$ defined by

$$L(t) = \frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\bigg|_{\mathcal{F}_t}. \tag{2.7.6}$$

Our aim is to show that this process has certain martingale properties, which can be extremely valuable in statistical applications. We shall also see that these properties only depend on the fact that $\mathbf{N}$ has compensator $\mathbf{\Lambda}$ or $\tilde{\mathbf{\Lambda}}$ under the two relevant probability measures, and that the special structure assumed in (2.7.1) is not needed. This is connected to the theory of *partial likelihood* which we discuss in the next section.

We can generalize and rewrite (2.7.2) as follows:

$$\begin{aligned}
L(t) &= L(0) \frac{\prod_{s \leq t} (\prod_h d\tilde{\Lambda}_h(s)^{\Delta N_h(s)} (1 - d\tilde{\Lambda}_\cdot(s))^{1 - \Delta N_\cdot(s)})}{\prod_{s \leq t} (\prod_h d\Lambda_h(s)^{\Delta N_h(s)} (1 - d\Lambda_\cdot(s))^{1 - \Delta N_\cdot(s)})} \\
&= L(0) \prod_{s \leq t} \left( \left( 1 - \frac{d\tilde{\Lambda}_\cdot(s) - d\Lambda_\cdot(s)}{1 - \Delta\Lambda_\cdot(s)} \right)^{1 - \Delta N_\cdot(s)} \right. \\
&\quad \left. \times \left( 1 - \sum_h \left( \frac{d\tilde{\Lambda}_h}{d\Lambda_h}(s) - 1 \right) dN_h(s) \right)^{\Delta N_\cdot(s)} \right) \\
&= L(0) \prod_{s \leq t} \left( 1 + \sum_h \left( \frac{d\tilde{\Lambda}_h}{d\Lambda_h}(s) - 1 - \frac{\Delta\tilde{\Lambda}_\cdot(s) - \Delta\Lambda_\cdot(s)}{1 - \Delta\Lambda_\cdot(s)} \right) \right. \\
&\quad \left. \times (dN_h(s) - d\Lambda_h(s)) \right),
\end{aligned} \tag{2.7.7}$$

where the last two steps follow by fomula (2.6.5) for the ratio of two real product-integrals and by some simple but tedious algebra.The derivation and the interpretation of (2.7.7) is not important; what is important is that by the Volterra integral equation characterization for product-integrals (Theorem II.6.1), $L$ is the unique solution of the equation

$$L(t) = L(0) + \sum_h \int_0^t L(s-)\left(\frac{d\tilde{\Lambda}_h}{d\Lambda_h}(s) - 1 - \frac{\Delta\tilde{\Lambda}_\cdot(s) - \Delta\Lambda_\cdot(s)}{1 - \Delta\Lambda_\cdot(s)}\right) dM_h(s), \tag{2.7.8}$$

where $M_h = N_h - \Lambda_h$, $h = 1, \ldots, k$, are $(P, (\mathscr{F}_t))$-local martingales. Because $L$ is cadlag and adapted, its left-continuous modification $L_-$ is predictable and locally bounded. If the predictable processes

$$\frac{d\tilde{\Lambda}_h}{d\Lambda_h} - 1 - \frac{\Delta\tilde{\Lambda}_\cdot - \Delta\Lambda_\cdot}{1 - \Delta\Lambda_\cdot}$$

are locally bounded too, we may conclude that $L$ itself—being the sum of integrals of locally bounded predictable processes with respect to local martingales—is a local martingale too.

In fact, one can show directly from (2.7.6) that $L$ is always a *martingale* when $\tilde{P} \ll P$. However, it is extremely useful that local martingale properties of the process

$$L(0)\frac{\prod_{s\le t}(\prod_h d\tilde{\Lambda}_h(s)^{\Delta N_h(s)}(1 - d\tilde{\Lambda}_\cdot(s))^{1-\Delta N_\cdot(s)})}{\prod_{s\le t}(\prod_h d\Lambda_h(s)^{\Delta N_h(s)}(1 - d\Lambda_\cdot(s))^{1-\Delta N_\cdot(s)})} \tag{2.7.9}$$

[and also with $L(0)$ deleted] can be derived quite independently of the interpretation, under (2.7.1) and $\tilde{P} \ll P$, of (2.7.9) as the likelihood ratio process $(L(t)) = ((d\tilde{P}/dP)|_{\mathscr{F}_t})$. In fact, local martingale properties of *score processes* and of *information processes* are also preserved, as we shall see later. These results are connected to the fact (which we explore further in the next section) that (2.7.9) can always be interpreted as the *partial likelihood* (ratio) based on observation of $\mathbf{N}$ and relative to some arbitrary filtration $(\mathscr{F}_t)$ not necessarily of the form $\mathscr{F}_t = \mathscr{F}_0 \vee \mathscr{N}_t$.

We finally sketch the related martingale properties of score and information processes. Suppose (2.7.1) holds for all $\{P_\theta: \theta \in \Theta\}$, for some open set $\Theta \subset \mathbb{R}$. Suppose all $P_\theta$ are dominated by a fixed probability measure, Q say. For simplicity, we assume that all $P_\theta$ coincide on $\mathscr{F}_0$ and work also in the absolutely continuous case: Under $P_\theta$, $\mathbf{N}$ has compensator $\Lambda^\theta = \int \lambda^\theta$ for certain intensity processes $\lambda^\theta$. We consider the likelihood function as depending on both $t \in \mathscr{T}$ and $\theta \in \Theta$; dropping the denominator in (2.7.4) (which does not depend on $\theta$), we have the likelihood at time $t$, as a function of $\theta$, is proportional to

$$L(\theta, t) = \exp(-\Lambda^\theta_\cdot(t)) \prod_{T_n \le t} \lambda^\theta_{J_n}(T_n) \tag{2.7.10}$$

and, hence,

$$\log L(\theta;t) = \sum_h \int_0^t \log \lambda_h^\theta(s)\,dN_h(s) - \Lambda_\cdot^\theta(t)$$
$$= \sum_h \int_0^t (\log \lambda_h^\theta(s)\,dN_h(s) - \lambda_h^\theta(s)\,ds).$$

To obtain the *score process*, we differentiate with respect to $\theta$. Supposing the derivative may be taken under the integral sign, we obtain

$$\frac{\partial}{\partial\theta}\log L(\theta;t) = \sum_h \int_0^t \frac{\partial}{\partial\theta}\log \lambda_h^\theta(s)(dN_h(s) - \lambda_h^\theta(s)\,ds), \tag{2.7.11}$$

so that again, under local boundedness or integrability conditions, the score process is a $(P_\theta,(\mathscr{F}_t))$-local martingale. Again, this holds without the assumptions leading to (2.7.10) being the likelihood functon. This result will be especially useful in Section VI.1 on maximum likelihood estimators.

Differentiating again with respect to $\theta$ to find the *observed information* at $\theta$, we find (if the differentiation can can again be taken under the integral sign)

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta;t) = \sum_h \int_0^t \frac{\partial^2}{\partial\theta^2}\log \lambda_h^\theta(s)\,dM_h(s) - \sum_h \int_0^t \left(\frac{\partial}{\partial\theta}\log \lambda_h^\theta(s)\right)^2 \lambda_h^\theta(s)\,ds. \tag{2.7.12}$$

Note that

$$\left\langle \frac{\partial}{\partial\theta}\log L(\theta;\cdot)\right\rangle = \sum_h \int \left(\frac{\partial}{\partial\theta}\log \lambda_h^\theta\right)^2 \lambda_h^\theta,$$

so that $\langle(\partial/\partial\theta)\log L(\theta;\cdot)\rangle$ is the compensator not only of $((\partial/\partial\theta)\log L(\theta;\cdot))^2$ but also of the process $-(\partial^2/\partial\theta^2)\log L(\theta;\cdot)$; this is a version of the well-known result: The variance of the score coincides with the expected information.

## II.7.3. Partial Likelihood

In Section II.7.1, we showed, cf. (2.7.2″), how the likelihood based on observation of a self-exciting multivariate counting process $\mathbf{N}$ with compensator $\mathbf{\Lambda}$ could be suggestively written in the form

$$dP = \prod_{t\in[0,\tau]} \left(\prod_h d\Lambda_h(t)^{dN_h(t)}(1 - d\Lambda_\cdot(t))^{1-dN_\cdot(t)}\right). \tag{2.7.13}$$

We also saw in Section II.7.2 that even if the process was not self-exciting—the filtration is larger than that generated by $\mathbf{N}$—then "likelihood ratios" and "score functions" calculated from the right-hand side of (2.7.13) still preserve various martingale properties, important in deriving statistical results for likelihood-based inference.

Here we link these results with the ideas of partial likelihood on the one hand (Cox, 1975; Kalbfleisch and Prentice, 1980) and with likelihoods for marked point processes on the other (Arjas and Haara, 1984; Arjas, 1989). The treatment will be heuristic and will belong more to statistical folklore than to mathematical probability theory. However, it is useful to discuss partial likelihood, even if only informally, in an abstract or general setting where the notation is not encumbered by specific details of particular applications. We will come to such applications in Chapter III; see, in particular, Section III.2.2.

The heuristic interpretation of (2.7.13) as a product of conditional multinomial probabilities ($n = 1$; $k + 1$ cells) connects naturally to the notion of *partial likelihood* (Cox, 1975). Suppose the data $X$ in a statistical problem can be represented as a sequence of smaller pieces of data $X_0, X_1, \ldots, X_n$. The density of $X$ can be factored as the product of conditional densities $p(x_i|x_0, \ldots, x_{i-1})$ of each $X_i$ given its predecessors $X_0, \ldots, X_{i-1}$. As a function of unknown parameters $\theta$ of the distribution of $X$, this is the likelihood for $\theta$ based on $X$. Cox (1975) claimed that if one deletes some of the factors, e.g., all the even numbered ($i = 0, 2, \ldots$), what remains (the product over $i = 1, 3, \ldots$) can still be used as a basis for statistical inference. In particular, the standard theory of large sample distributional properties of maximum likelihood estimators, likelihood ratio tests and score tests, will also apply to the *partial likelihood.*

The reason for this is actually martingale based, though Cox (1975) did not explicitly refer to martingales in his paper. The components of the partial score $(\partial/\partial\theta) \log p(X_i|X_0, X_1, \ldots, X_{i-1}; \theta)$, $i = 1, 3, \ldots$, form a so-called discrete-time martingale difference sequence and, from this, it follows that the log partial likelihood satisfies the usual relations between expectations of first and second derivatives. Moreover, if the number of components is large and their individual influence small, the martingale central limit theorem could supply asymptotic normality of the partial score.

This suggests that we could also delete terms from (2.7.13) as desired and use what remains as a partial likelihood. If one deletes the contributions for the time interval $dt$ according to the value of a predictable indicator process $C$, the result is called "the partial likelihood for the filtered counting process $\int C\, d\mathbf{N}$," to which we return in Section III.4.

It is more useful to base partial likelihoods on further factorizations of (2.7.13). Suppose the process $\mathbf{N}$ registers two distinct kinds of events, which may even occur simultaneously. Typically, these will be *failure events* and *censoring* or *covariate* events. We can correspondingly factor each multinomial likelihood in (2.7.13) into a likelihood based on the marginal distribution of the first type of event and the likelihood based on the conditional distribution of the second given the first. Then we could form a partial likelihood by keeping only the first (marginal) pieces of each pair corresponding to the failure events. Such a partial likelihood (based on failure events, disregarding censoring and covariate events) is often used in sur-

vival analysis; see especially Sections III.2 and III.5 of this volume. One could also apply the idea to classification of the events into more than two levels, keeping just the contribution from one particular level conditional on the previous levels. The famous Cox partial likelihood (Example VII.2.1) used to derive the usual estimator in the Cox regression model (Cox, 1972) is of this type: One factors according to whether or not there is a failure event; if so, then to which individual it occurs, and then, given the preceding, one looks at possible censorings or changes in covariates. The Cox partial likelihood corresponds to keeping just the middle term of each triple (which individual failed, given there was a failure); one disregards the first term of each triple saying whether or not there is a failure, and the third term, saying what censoring or covariate events took place. In both these examples, the partial likelihood makes inference easier by discarding factors which depend in a complicated or even unknown way on nuisance parameters.

To formalize these ideas, we can use the fact that a classification of the events registered by $\mathbf{N}$ corresponds to aggregation or grouping of some or all of the $k+1$ multinomial cells in (2.7.13) to a smaller number of cells. Multinomial probabilities will then be written as products of marginal (multinomial) probabilities for the coarse classification of events and conditional probabilities for the fine classification given the first one. There is no difficulty at all in extending this to a hierarchy of classifications. A partial likelihood is formed by collecting just the contributions from one particular level of the hierarchy of classifications.

It is convenient at this point to generalize to marked point process notation (see Section II.4.1); the ideas are not changed though some technical details become more delicate through the added generality. In fact, we will only consider countable mark spaces in the rest of this book, but the restriction does not lead to substantial simplification.

The likelihood in the self-exciting case (2.7.13) is now written as

$$\prod_t \left( \prod_{x \in E} \Lambda(\mathrm{d}t, \mathrm{d}x)^{\mathbf{N}(\mathrm{d}t, \mathrm{d}x)} (1 - \Lambda(\mathrm{d}t, E))^{1 - \mathbf{N}(\mathrm{d}t, E)} \right), \tag{2.7.14}$$

where $\mathbf{N}$ is a marked point process with marks in $(E, \mathscr{E})$ and $\Lambda$ is its compensator; $\mathbf{N}$ and $\Lambda$ are considered as (random) measures on $(\mathscr{T} \times E, \mathscr{B}(\mathscr{T}) \otimes \mathscr{E})$. The interpretation of (2.7.14) remains the same: Conditional on the past at time $t$, there is an event in the time interval $[t, t + \mathrm{d}t)$, or just $\mathrm{d}t$ for short, with mark in $\mathrm{d}x$ [so $\mathbf{N}(\mathrm{d}t, \mathrm{d}x) = 1$] with probability $\Lambda(\mathrm{d}t, \mathrm{d}x)$; there is no event [$\mathbf{N}(\mathrm{d}t, E) = 0$] with probability $1 - \Lambda(\mathrm{d}t, E)$. Likelihood *ratios* are formed, as explained in the discussion following Theorem II.7.2, by collecting the finite number of contributions to (2.7.14) where $\mathbf{N}(\mathrm{d}t, \mathrm{d}x) = 1$ and forming Radon–Nikodym derivatives $\Lambda(\mathrm{d}t, \mathrm{d}x)/\tilde{\Lambda}(\mathrm{d}t, \mathrm{d}x) = (\mathrm{d}\Lambda/\mathrm{d}\tilde{\Lambda})(t, x)$. The remaining contributions are standard product-integrals over $\mathscr{T}$ less the finite number of time points where an event occurs.

We need a convenient notation for a hierarchy of classifications of type of event, the finest level being provided by $(E, \mathscr{E})$ itself. Arjas (1989) proposed

that such classifications be regarded as a function $g$ on the mark space $E$ to a (typically smaller) space $G$, so $g$ groups or aggregates the marks in $E$, putting them together into certain classes characterized by the value of $g$. This is called the *premark approach* with the idea that the occurrence of the premark is some kind of signal that more is going to come. A complication arises because we may want to group some of the events in $E$ with the "nonevent," thus discarding some of the time points at which events occur altogether. Thus, the premark could indicate failures; the full mark adds to the description of what failures occur at a given time also what censorings occur. It is possible for the premark to be empty while the full mark does indicate that an event has taken place (censoring without simultaneous failure). Rather than the term premark, we will use the more neutral and accurate term *reduced mark*. Another common terminology is to talk of *innovative* and *noninnovative* events (or informative and noninformative): The innovative events are the occurrences of the premarks; the noninnovative events are whatever else can happen.

We show how a function $g$ on the mark space reduces or aggregates $\mathbf{N}$ to a "smaller" marked point process $\mathbf{N}^g$ with reduced mark space $G$. The compensator $\mathbf{\Lambda}^g$ of $\mathbf{N}^g$ will be easily computable from $\mathbf{\Lambda}$, the compensator of $\mathbf{N}$. Then we use the multinomial analogy to suggest how (2.7.14) would factor under this aggregation. A complication comes from having to deal with "no event" (both for $\mathbf{N}$ and for $\mathbf{N}^g$) explicitly. We then write down the analogous factorization of (2.7.14) and show how each of the terms arising can be given a formal (measure-theoretic) interpretation, taking account of the fact that the final expression should make mathematical sense on forming a ratio with a similar expression under a different probability measure (same $\mathbf{N}$, different compensator $\tilde{\mathbf{\Lambda}}$). So the approach is a heuristic one which leads us to a factorization of (2.7.14) whose validity can be established by a direct (tedious) calculation. The approach also leads to interesting conjectures: If the conditional terms which we would like to delete to form a partial likelihood do *not* depend on the parameter of interest anyway, then the result is not just a *partial* likelihood but actually the *full* likelihood for this parameter.

Let $\mathbf{N}$ be a self-exciting marked point process with compensator $\mathbf{\Lambda}$ and mark space $E$. Let $\varnothing$ be a point *not* in $E$ which we call the "empty mark," and use to represent "no event." Let $\bar{E} = E \cup \{\varnothing\}$. Let $G$ be another mark space, suppose $\varnothing \notin G$, and let $\bar{G} = G \cup \{\varnothing\}$. (Alternatively one can use a different symbol for the empty mark in each space $E$ and $G$, but this makes things look more complicated than necessary.) Let $g\colon \bar{E} \to \bar{G}$ be a measurable mapping such that $g(\varnothing) = \varnothing$, "the empty mark in $E$" is mapped to "the empty mark in $G$." Typically, $g$ will be a many-to-one mapping. ($E$ and $G$ are equipped with $\sigma$-algebras $\mathscr{E}$ and $\mathscr{G}$; $\bar{E}$ and $\bar{G}$ are given the $\sigma$-algebras on $\bar{E}$, $\bar{G}$, *generated* by $\mathscr{E}$ and $\mathscr{G}$.)

Let $\mathbf{N}^g$ be the marked point process with mark space $G$ defined by $\mathbf{N}^g((0,t] \times A) = \mathbf{N}((0,t] \times g^{-1}(A))$. Thus, $\mathbf{N}^g$ as a point process has points $(T_n, g(J_n))$ for those points $(T_n, J_n)$ of $\mathbf{N}$ such that $g(J_n) \neq \varnothing$. The compensator

of $\mathbf{N}^g$ is $\Lambda^g$ defined by

$$\Lambda^g((0,t] \times A) = \Lambda((0,t] \times g^{-1}(A)).$$

The conditional distribution, given $\mathcal{F}_{t-}$ of the events in $dt$, can now be built up as follows:

with probability $\Lambda^g(dt, dy)$, the reduced process has an event in $dt \times dy$, i.e., $\mathbf{N}^g(dt, dy) = 1$; with probability $1 - \Lambda^g(dt, G)$, it has no event, i.e., $\mathbf{N}^g(dt, G) = 0$;

*given* the reduced process has an event in $dt \times dy$, i.e., $\mathbf{N}^g(dt, dy) = 1$, the original process has one in $dt \times dx$, i.e., $\mathbf{N}(dt, dx) = 1$ [for $x$ with $g(x) = y$], with conditional probability $\Lambda(dt, dx)/\Lambda^g(dt, dy)$;

but *given* the reduced process has *no* event in $dt \times dy$, i.e., $\mathbf{N}^g(dt, G) = 0$, the original process has an event in $dt \times dx$, i.e., $\mathbf{N}(dt, dx) = 1$ [for $x$ with $g(x) = \varnothing$], with conditional probability $\Lambda(dt, dx)/(1 - \Lambda^g(dt, G))$; the original process has no event, i.e., $\mathbf{N}(dt, E) = 0$, with the complementary conditional probability $1 - \Lambda(dt, g^{-1}(\varnothing))/(1 - \Lambda^g(dt, G))$.

Combining all these possibilities in the same order as we have just described them suggests that (2.7.14) can be rewritten as

$$\begin{aligned} d\mathbf{P} = \prod_t \Bigg\{ & \Bigg( \prod_{y \in G} \Lambda^g(dt, dy)^{\mathbf{N}^g(dt, dy)} (1 - \Lambda^g(dt, G))^{1 - \mathbf{N}^g(dt, G)} \Bigg) \\ & \cdot \prod_{y \in G} \Bigg( \prod_{x: g(x) = y} \left( \frac{\Lambda(dt, dx)}{\Lambda^g(dt, dy)} \right)^{\mathbf{N}(dt, dx)} \Bigg)^{\mathbf{N}^g(dt, dy)} \\ & \cdot \Bigg( \prod_{x: g(x) = \varnothing} \left( \frac{\Lambda(dt, dx)}{1 - \Lambda^g(dt, G)} \right)^{\mathbf{N}(dt, dx)} \\ & \cdot \left( 1 - \frac{\Lambda(dt, g^{-1}(\varnothing))}{1 - \Lambda^g(dt, G)} \right)^{1 - \mathbf{N}(dt, E)} \Bigg)^{1 - \mathbf{N}^g(dt, G)} \Bigg\}. \end{aligned} \tag{2.7.15}$$

The intuitive and probabilistic interpretation of (2.7.15) is not really difficult, even if the formula is rather long. A rigorous mathematical interpretation is, however, a little delicate.

The first line of (2.7.15) can be mathematically interpreted as it stands (on taking a ratio with $d\tilde{\mathbf{P}}$, forming Radon–Nikodym derivatives and product-integrals) and gives us the *partial likelihood* based on $\mathbf{N}^g$ (with the rest of the information in $\mathbf{N}$ ignored). Note that this partial likelihood has exactly the same form as the likelihood based on $\mathbf{N}^g$ alone in the case that this process is self-exciting; however, now $\Lambda^g$ may depend on the whole past of $\mathbf{N}$, not just on $\mathbf{N}^g$. The second and last two lines provide the other part of the factorization and are harder to interpret mathematically. We look at each part in turn.

The second line gives a contribution $\Lambda(dt, dx)/\Lambda^g(dt, dy)$ for $t$, $x$, and $y = g(x)$ which corresponds to events of both $\mathbf{N}$ and $\mathbf{N}^g$, a finite number of terms. Because $\Lambda^g$ is a "marginalization" of $\Lambda$, one can "disintegrate" $\Lambda$ [restricted to $\mathcal{T} \times E \backslash g^{-1}(\varnothing)$] into the product of the image $\Lambda^g$ of $\Lambda$ and a *transition probability measure* $\Lambda(dx|t, y)$ on $\{x: g(x) = y\}$; thus, one may write

$$\Lambda(\mathrm{d}t, \mathrm{d}x) = \Lambda^g(\mathrm{d}t, \mathrm{d}y)\Lambda(\mathrm{d}x|t, y)$$

in the sense that an integral over $t$ and $x$ on the left-hand side equals a triple integral over $t$, $y$, and $x = g(y)$ on the right-hand side. Intuitively, for $y = g(x)$, we write the probability of a mark in $\mathrm{d}x$ and time $\mathrm{d}t$, given the past, as the probability of a reduced mark in $\mathrm{d}y$ and time $\mathrm{d}t$, given the past, times the probability of a mark in $\mathrm{d}x$ given a reduced mark $y$ at time $t$. So $\Lambda(\mathrm{d}t, \mathrm{d}x)/\Lambda^g(\mathrm{d}t, \mathrm{d}y)$ has a mathematical interpretation as $\Lambda(\mathrm{d}x|t, y)$, and the ratio of two such terms (for compensators $\Lambda$ and $\tilde{\Lambda}$ under two different probability measures P and $\tilde{\mathrm{P}}$) can be mathematically interpreted as the Radon–Nikodym derivative, for the given point $t$, $y$, of the transition probability measure $\Lambda(\cdot\,|t, y)$ with respect to $\tilde{\Lambda}(\cdot\,|t, y)$ on $\{x\colon g(x) = y\}$.

The third line also only occurs in a finite number of terms. This suggests mathematically interpreting $\Lambda(\mathrm{d}t, \mathrm{d}x)/(1 - \Lambda^g(\mathrm{d}t, G))$ as $(1 - \Lambda^g(\{t\} \times G))^{-1} \cdot \Lambda(\mathrm{d}t, \mathrm{d}x)$, ratios of which can be mathematically interpreted as a Radon–Nikodym derivative $(\mathrm{d}\Lambda/\mathrm{d}\tilde{\Lambda})(t, x)$ times the function $(1 - \tilde{\Lambda}^g(\{t\} \times G))/(1 - \Lambda^g(\{t\} \times G))$.

The fourth line is present for all $t$ such that $\mathbf{N}$ has no event in $\mathrm{d}t$. Because

$$1 - \frac{\Lambda(\mathrm{d}t, g^{-1}(\varnothing))}{1 - \Lambda^g(\mathrm{d}t, G)} = \frac{1 - \Lambda(\mathrm{d}t, g^{-1}(\varnothing)) - \Lambda^g(\mathrm{d}t, G)}{1 - \Lambda^g(\mathrm{d}t, G)} = \frac{1 - \Lambda(\mathrm{d}t, E)}{1 - \Lambda^g(\mathrm{d}t, G)},$$

we can mathematically interpret a product over $t$ of terms like this as a ratio of product-integrals.

To sum up, we may rewrite (2.7.15) as

$$\begin{aligned}\mathrm{dP} = \prod_t \Bigg\{ &\Bigg( \prod_{y \in G} \Lambda^g(\mathrm{d}t, \mathrm{d}y)^{\mathbf{N}^g(\mathrm{d}t, \mathrm{d}y)}(1 - \Lambda^g(\mathrm{d}t, G))^{1 - \mathbf{N}^g(\mathrm{d}t, G)} \Bigg) \\ &\cdot \prod_{x\colon g(x) \neq \varnothing} \Lambda(\mathrm{d}x|t, g(x))^{\mathbf{N}(\mathrm{d}t, \mathrm{d}x)} \prod_{x\colon g(x) = \varnothing} \left( \frac{\Lambda(\mathrm{d}t, \mathrm{d}x)}{1 - \Lambda^g(\{t\} \times G)} \right)^{\mathbf{N}(\mathrm{d}t, \mathrm{d}x)} \\ &\cdot \left( \frac{1 - \Lambda(\mathrm{d}t, E)}{1 - \Lambda^g(\mathrm{d}t, G)} \right)^{1 - \mathbf{N}(\mathrm{d}t, E)} \Bigg\}, \end{aligned} \tag{2.7.16}$$

which may be mathematically interpreted by taking ratios and forming Radon–Nikodym derivatives and product-integrals. The first line is the partial likelihood based on $\mathbf{N}^g$; the next two lines (together) form a partial likelihood based on the rest of $\mathbf{N}$.

A special case is when $E$ and $G$ are countable and $\Lambda$ is absolutely continuous on $\mathscr{T} \times E$ with respect to Lebesgue measure times counting measure. This means that $\mathbf{N}$ is a counting process in the usual sense (with a countable number of components) and $\mathbf{N}^g$ is some aggregration of $\mathbf{N}$; both have intensities, say $\lambda_x$: $x \in E$ and $\lambda_y^g$, $y \in G$. Then $\lambda_y^g(t) = \sum_{x\colon g(x) = y} \lambda_x(t)$; the transition measure $\Lambda(\mathrm{d}x|t, g(x))$ is a probability measure on the finite set $\{x\colon g(x) = y\}$ and its atoms are simply $\lambda_x(t)/\lambda_y^g(t)$. The atomic part $1 - \Lambda^g(\{t\} \times G)$ disappears and the factorization is

$$
\begin{aligned}
\mathrm{dP} \propto \prod_t \Bigg\{ & \Bigg( \prod_y \lambda_y^g(t)^{N_y^g(\mathrm{d}t)} (1 - \lambda_\bullet^g(t)\,\mathrm{d}t)^{1 - N_\bullet^g(\mathrm{d}t)} \Bigg) \\
& \cdot \prod_{x:\, g(x) \neq \varnothing} \left( \frac{\lambda_x(t)}{\lambda_{g(x)}^g(t)} \right)^{N_x(\mathrm{d}t)} \prod_{x:\, g(x) = \varnothing} \lambda_x(t)^{N_x(\mathrm{d}t)} \left( \frac{1 - \lambda_\bullet(t)\,\mathrm{d}t}{1 - \lambda_\bullet^g(t)\,\mathrm{d}t} \right)^{1 - N_\bullet(\mathrm{d}t)} \Bigg\} \\
\propto & \prod_{t,y} \lambda_y^g(t)^{N_y^g(\mathrm{d}t)} \exp\left( - \int_0^\tau \lambda_\bullet^g(t)\,\mathrm{d}t \right) \prod_{t,x:\, g(x) \neq \varnothing} \left( \frac{\lambda_x(t)}{\lambda_{g(x)}^y(t)} \right)^{N_x(\mathrm{d}t)} \\
& \cdot \prod_{t,x:\, g(x) = \varnothing} \lambda_x(t)^{N_x(\mathrm{d}t)} \exp\left( - \int_0^\tau (\lambda_\bullet(t) - \lambda_\bullet^g(t))\,\mathrm{d}t \right).
\end{aligned}
$$

One easily sees that this really is a factorization of the full likelihood

$$
\mathrm{dP} \propto \prod_t \left( \prod_x \lambda_x(t)^{N_x(\mathrm{d}t)} (1 - \lambda_\bullet(t)\,\mathrm{d}t)^{1 - N_\bullet(\mathrm{d}t)} \right),
$$

as must, of course, be the case.

It is possible (see the next paragraph) to give conditions under which the factorization (2.7.16) is correct; probably it is also possible to show that the two factors of (2.7.16) (the first line versus the last three) can be obtained as limits of proper, finite, partial likelihoods in the sense of Cox (1975); see the results of Slud (1991). However, such mathematical results, though supporting the intuition, do not lead to any useful mathematical properties of partial likelihood which cannot be obtained easily and directly. So, it is quite justified to leave the matter in its present informal state: We have described a valuable heuristic tool, not presented a formal mathematical theory.

The following remarks form a technical aside on the regularity conditions on the mark spaces $E$ and $G$ needed to justify all the steps here. Arjas and Haara (1984) assumed that $E$ is the product of a countable set with a Polish space, and $g$ is the coordinate projection onto the first coordinate. An interesting generalization due to Arjas (1989) would be to let the function $g$ be replaced by a predictable process, so that classification of events may change randomly (but predictably) in time. We do not, however, follow up this idea here. Arjas and Haara (1984) used a representation of $E$ as a product space (innovative versus noninnovative marks), with the classification of events seen as a *projection* on the first coordinate (aggregation over the second). Again, the need to consider "no event" leads to an involved notation. The reduced mark approach and the product mark space approach are mathematically equivalent.

## II.8 The Functional Delta-Method

The delta-method is a popular and elementary tool of asymptotic statistics. One can summarize it in the following statements: Suppose for some random $p$-vectors $\mathbf{T}_n$ and a sequence of numbers $a_n \to \infty$,

$$a_n(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathbf{Z} \quad \text{as } n \to \infty,$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is fixed. Suppose $\boldsymbol{\phi}: \mathbb{R}^p \to \mathbb{R}^q$ is differentiable at $\boldsymbol{\theta}$ with $q \times p$ matrix $\boldsymbol{\phi}'(\boldsymbol{\theta})$ of partial derivatives. Then

$$a_n(\boldsymbol{\phi}(\mathbf{T}_n) - \boldsymbol{\phi}(\boldsymbol{\theta})) \xrightarrow{\mathcal{D}} \boldsymbol{\phi}'(\boldsymbol{\theta}) \cdot \mathbf{Z} \quad \text{in } \mathbb{R}^q$$

and, indeed, $a_n(\boldsymbol{\phi}(\mathbf{T}_n) - \boldsymbol{\phi}(\boldsymbol{\theta}))$ is asymptotically equivalent to $\boldsymbol{\phi}'(\boldsymbol{\theta}) \cdot a_n(\mathbf{T}_n - \boldsymbol{\theta})$.

Our statistical models will frequently be non- or semiparametric models, and we will estimate cumulative hazard functions, survival functions, and the like, getting their large sample properties by application of the martingale central limit theorem. However, we will often want to transfer asymptotic normality to various functionals of interest (cf. Section IV.3.4). Also, in Section VIII.3 we need to transfer asymptotic efficiency from estimators to functionals of estimators. All this can be done using an infinite-dimensional version of the delta-method.

This indeed exists, but to make its use more elegant, we will need to switch from the usual weak convergence theory for $D[0, \tau]$ based on the Skorohod metric (Billingsley, 1968) to a perhaps less familiar theory due to Dudley (1966) and popularized by Pollard (1984) based on the supremum norm. The $\sigma$-algebra on $D[0, \tau]$ used in this theory is exactly the same, but is now interpreted as the $\sigma$-algebra generated by the supremum-norm open balls in $D[0, \tau]$, or equivalently as the $\sigma$-algebra generated by the *coordinate mappings* $x \mapsto x(t)$ mapping $D[0, \tau]$ to $\mathbb{R}$. (These coincidences perhaps explain the great usefulness of the otherwise rather complicated Skorohod metric.) The open-ball $\sigma$-algebra is *strictly smaller* than the (supremum-norm) Borel $\sigma$-algebra; this is related to the nonseparability of $D[0, \tau]$ under the supremum norm.

Weak convergence in $D[0, \tau]$, supremum norm, now means convergence of the expectations of all bounded, real, continuous *measurable* functions of the random element concerned. If the limiting process has continuous sample paths, weak convergence in the sense of the Skohorod metric and in the sense of the supremum norm are exactly equivalent. Otherwise, sup-norm convergence is *stronger*.

We will consider random elements of spaces like $B = D[0, \tau]^p \times \mathbb{R}^q$ endowed with, e.g., the max-supremum norm, or another convenient equivalent norm. This makes $B$ a Banach space (not necessarily separable). Weak convergence of random elements of $B$ has already been described; we now need to introduce a concept of *differentiability* suitable for statistical applications, allowing us to generalize the delta-method.

This is provided by the concept of *Hadamard* or *compact* differentiability, intermediate between the more familiar but too strong notion of *Fréchet* or *bounded* differentiability, and the too weak *Gâteaux* or *directional* differentiability. See Gill (1989) for a more extensive introduction to these matters, on which this summary is based.

Many equivalent formulations of Hadamard differentiability exist. Most convenient for us is the following definition:

**Definition II.8.1.** Let $B$, $B'$ be spaces of the type just described. $\phi: B \to B'$ is compactly or Hadamard differentiable at a point $\theta \in B$ if and only if a continuous, linear map

$$d\phi(\theta): B \to B'$$

exists (called the derivative of $\phi$ at the point $\theta$) such that for all real sequences $a_n \to \infty$ and all convergent sequences $h_n \to h \in B$,

$$a_n(\phi(\theta + a_n^{-1}h_n) - \phi(\theta)) \to d\phi(\theta) \cdot h \quad \text{as } n \to \infty. \tag{2.8.1}$$

The notation $d\phi(\theta) \cdot h$ for the linear mapping $d\phi(\theta)$ acting on the element $h \in B$ is supposed to suggest multiplication. When $B$ and $B'$ are Euclidean and $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are real vectors, $d\phi(\theta)$ as a linear map from one Euclidean space to another can be represented by a matrix multiplication, and $d\phi(\theta)$ can be identified with the matrix of partial derivatives of the components of $\boldsymbol{\phi}$ with respect to those of $\boldsymbol{\theta}$.

This allows us to give a functional version of the delta-method.

**Theorem II.8.1.** *Let $T_n$ be a sequence of random elements of $B$, $a_n \to \infty$ a real sequence, such that*

$$a_n(T_n - \theta) \xrightarrow{\mathcal{D}} Z$$

*for some fixed point $\theta \in B$ and a random element $Z$ of $B$. Suppose $\phi: B \to B'$ is compactly differentiable at $\theta$. Then*

$$a_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{D}} d\phi(\theta) \cdot Z$$

*and, moreover,*

$$a_n(\phi(T_n) - \phi(\theta)) \quad \text{and} \quad d\phi(\theta) \cdot a_n(T_n - \theta)$$

*are asymptotically equivalent.*

The proof of Theorem II.8.1 is extremely simple. One invokes the so-called Skorohod–Dudley almost sure representation theorem (see, e.g., Pollard, 1984 or 1990) to establish [because of the weak convergence of $a_n(T_n - \theta)$ to $Z$] the existence of versions of $T_n$ and of $Z$, all defined on a single probability space, and such that $a_n(T_n - \theta)$ converges *almost surely* to $Z$. Now one applies the definition of compact differentiability to show that, on this new probability space, $a_n(\phi(T_n) - \phi(\theta))$ converges almost surely to $d\phi(\theta) \cdot Z$. Almost sure convergence implies convergence in distribution, but the new and the old versions of $a_n(\phi(T_n) - \phi(\theta))$ have exactly the same distributions for each $n$ by the Skorohod–Dudley construction. Hence, we get the required result.

Sometimes, (2.8.1) is not satisfied for all sequences, but only for sequences $h_n \in B$ with limit $h$ in a subset $H \subseteq B$. If the limiting process $Z$ has sample paths in $H$, then this is still enough for the conclusions of the delta-method. We, therefore, make the following definition:

**Definition II.8.2.** Let $B$, $B'$, $\phi$, $\theta$, and $d\phi(\theta)$ be as in Definition II.8.1, and suppose $H \subseteq B$ exists such that for all sequences $h_n \in B$ with $h_n \to h \in H$, (2.8.1) holds. Then we say that $\phi$ is tangentially compactly differentiable at $\theta$ (tangentially to $H$).

Then we have the delta-method again:

**Theorem II.8.2.** *Let $T_n$, $B$, $\phi$, $\theta$, and $Z$ be as in Theorem* II.8.1, *except that $\phi$ is only differentiable tangential to $H$, but $Z$ has sample paths in $H$. Then the conclusions of Theorem* II.8.1 *still remain true.*

Sometimes a functional $\phi$ is only defined on some subset $E \subseteq B$ (maybe, $T_n$ only takes values in $E$, so the definition of $\phi$ outside $E$ should be arbitrary). It turns out that one need only verify (2.8.1) for $\theta + a_n^{-1}h_n \in E$ (for either proper or only tangential differentiability): $\phi$ can then be extended to $B$ so as to be (tangentially) differentiable at $\theta$:

**Lemma II.8.3.** *Let $E \subseteq B$ and let $\phi$: $E \to B'$ be given. Suppose $\theta \in E$ and* (2.8.1) *holds for all sequences $a_n$, $h_n$ such that $\theta + a_n^{-1}h_n \in E$ for all $n$ and $h_n \to h$ ($\in H$, in the tangential case). Then $\phi$ can be defined on all of $B$ so as to be (tangentially) differentiable at $\theta$ with derivative $d\phi(\theta)$.*

Before we move to applications, here is a most important remark. Compact differentiability (also tangential differentiability, under the obvious compatibility conditions) satisfies the *chain rule*: The composition of differentiable functions is differentiable, with derivative equal to the composition (product) of the derivatives; or

$$d(\psi \circ \phi)(\theta) = d\psi(\phi(\theta)) \cdot d\phi(\theta).$$

This means that from the differentiability of the few basic functionals in what follows, the differentiability of a large class of composed functionals follows too without any further work. As we saw in the sketch of the proof of Theorem II.8.1, the functional delta-method is really no more than a way to package the Skorohod–Dudley almost sure convergence theorem. The chain rule further allows us to modularize its application: We decompose a complicated functional into a composition of simple and perhaps familiar functionals whose differentiability is known or easy to check.

Now for applications. The most common are quantiles and inverses, integrals and product-integrals, and composition. We mean here a different sort of composition to that referred to just now: Take two elements $x$ and $y$ of $D[0,\tau]$ such that $y$ also takes values in $[0,\tau]$ and form the new element $x(y(\cdot)) \in D[0,\tau]$. All these mappings are (tangentially) differentiable at points satisfying natural restrictions, and so therefore are compositions of these mappings (e.g., the inverse of the product-integral of the ordinary integral of one element of $D[0,\tau]$ with respect to another).

**Proposition II.8.4** (The $p$th Quantile). *Let $E$ consist of all nondecreasing elements of $B = D[0,\tau]$, and let $\phi(x) = x^{-1}(p) = \inf\{t: x(t) \geq p\}$ for a fixed $p \in \mathbb{R}$. Suppose $\theta \in E$ is such that $0 < \theta^{-1}(p) < \tau$ and $\theta$ is ordinarily differentiable at $\theta^{-1}(p)$. Then $\phi$ is tangentially differentiable at $\theta$, taking $H = \{h \in B$: $h$ is continuous at $\theta^{-1}(p)\}$, with derivative*

$$d\phi(\theta)\cdot h = -\frac{h(\theta^{-1}(p))}{\theta'(\theta^{-1}(p))}.$$

**Proposition II.8.5** (The Inverse). *Let $E$ be as in Proposition* II.8.4 *and let $\phi(x) = x^{-1}$ restricted to a fixed interval $[p_0, p_1]$. Suppose $\theta \in E$ is such that $\theta$ is (ordinarily) continuously differentiable on $[\theta^{-1}(p_0) - \varepsilon, \theta^{-1}(p_1) + \varepsilon]$ with positive derivative there, for some $\varepsilon > 0$. Then $\phi$ is tangentially differentiable at $\theta$, taking $H$ to be $C[0,\tau]$, the subset of continuous functions. The derivative is given by*

$$d\phi(\theta)\cdot h = -\left(\frac{h}{\theta'}\right) \circ \theta^{-1}$$

*restricted to $[p_0, p_1]$.*

In both of the preceding propositions, $\phi$ was only defined on a subset of $D[0,\tau]$, but Lemma II.8.3 is applicable to extend it in some way to all of $D[0,\tau]$. A similar remark must be made for all our following examples.

**Proposition II.8.6** (Integration). *Let $E = D[0,\tau] \times E_M$ where $E_M \subseteq D[0,\tau]$ is the set of functions of total variation bounded by the constant $M$. Let*

$$\phi: E \to D[0,\tau]$$

*be defined by $\phi(x,y) = \int x\,dy$. Then $\phi$ is compactly differentiable at a point $(x,y)$ of $E$ such that $x$ is of bounded variation too, with*

$$d\phi(x,y)\cdot(h,k) = \int h\,dy + \int x\,dk,$$

*where the last integral is defined by the integration by parts formula ($k$ may not be of bounded variation)*

$$\int x\,dk = xk - x(0)k(0) - \int k_-\,dx.$$

A similar *continuity* property of this mapping holds under less stringent conditions: If $(x_n, y_n) \in E$ is a sequence converging (in supremum norm) to $(x,y) \in E$, where $x$ need not be of bounded variation, then $\int x_n\,dy_n \to \int x\,dy$ in supremum norm. This result is a version of the classical *Helly–Bray lemma*; see Gill (1989, Lemma 3 and subsequent discussion).

The following result on product-integration (Definition II.6.1) is actually a consequence of the Duhamel equation (2.6.4).

**Proposition II.8.7** (Product-Integration). *Let* $E_M^{k^2} \subseteq (D[0,\tau])^{k^2}$ *be the set of* $k \times k$ *matrix cadlag functions with components of total variation bounded by the constant* $M$. *Let* $\phi\colon E_M^{k^2} \to (D[0,\tau])^{k^2}$ *be defined by*

$$\phi(\mathbf{X}) = \prod (\mathbf{I} + \mathrm{d}\mathbf{X}).$$

*Then* $\phi$ *is compactly differentiable at each point of* $E_M^{k^2}$ *with derivative* $\mathrm{d}\phi(\mathbf{X})\cdot\mathbf{H}$ *given by*

$$(\mathrm{d}\phi(\mathbf{X})\cdot\mathbf{H})(t) = \int_{s\in[0,t]} \prod_{[0,s)} (\mathbf{I} + \mathrm{d}\mathbf{X})\mathbf{H}(\mathrm{d}s) \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{X}), \tag{2.8.2}$$

*where the last integral is defined by application (twice) of the integration by parts formula.*

For completeness, the integration by parts formula applied to $\int \prod (\mathbf{I} + \mathrm{d}\mathbf{X})\,\mathrm{d}\mathbf{Z}$ and to $\mathbf{Z} = \int \mathrm{d}\mathbf{H} \prod (\mathbf{I} + \mathrm{d}\mathbf{X})$ gives

$$\begin{aligned}(\mathrm{d}\phi(\mathbf{X})\cdot\mathbf{H})(t) &= \prod_{(0,t]} (\mathbf{I} + \mathrm{d}\mathbf{X})\mathbf{H}(t) \\ &\quad + \int_{s\in[0,t]} \prod_{[0,s)} (\mathbf{I} + \mathrm{d}\mathbf{X})(\mathbf{X}(\mathrm{d}s)\mathbf{H}(s) - \mathbf{H}(s)\mathbf{X}(\mathrm{d}s)) \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{X}),\end{aligned}$$

showing that $\mathrm{d}\phi(\mathbf{X})$ really is a continuous linear map from $(D[0,\tau])^{k^2}$ to itself. A version of this formula can be found in Fleming (1978b).

One can also obtain a simple continuity result from the Duhamel equation: If $\mathbf{X}_n$, $\mathbf{X}$ in $E_M^{k^2}$ are such that $\mathbf{X}_n \to \mathbf{X}$ in supremum norm, then $\prod (\mathbf{I} + \mathrm{d}\mathbf{X}_n) \to \prod (\mathbf{I} + \mathrm{d}\mathbf{X})$ in supremum norm too.

Finally, a result on composition:

**Proposition II.8.8** (Composition). *Let* $E \subseteq (D[0,\tau])^2$ *be the set of nondecreasing functions on* $[0,\tau]$ *and let* $\theta = (x,y)$ *be a fixed point in* $E$ *such that* $0 < y(0) \le y(\tau) < \tau$ *and such that* $x$ *is (ordinarily) continuously differentiable on* $[0,\tau]$. *Then the mapping* $(x,y) \mapsto x \circ y$ *from* $E$ *to* $D[0,\tau]$ *is compactly differentiable tangentially to* $C[0,\tau] \times D[0,\tau]$ *at* $\theta$ *with derivative*

$$\mathrm{d}\phi(x,y)\cdot(h,k) = h \circ y + x' \circ y \cdot k.$$

The composition and inverse mappings occur when plotting one empirical function versus another as in a P-P plot for instance; a plot of $Y(t)$ against $X(t)$ for $t$ in some interval is actually a plot of $Y(X^{-1}(p))$ against $p$ for certain $p$. Product-integration occurs when computing survival functions from hazard functions. One can imagine elaborate examples such as the quantile residual lifetime function seen as a functional of the cumulative hazard function. (The residual lifetime distribution of a survival time $T$ is actually a family of distributions, namely, of the residual lifetime $T - t$ conditional on $T > t$ for each $t > 0$.)

# II.9. Bibliographic Remarks

Martingale theory has a long history and plays a fundamental role in all parts of modern probability theory. A first appearance seems to have been in the paper by Bernstein (1926) under the name *variables enchaînées*. In fact, this paper is devoted to central limit theory and, in particular, its generalization to sums of *dependent* random variables; in particular, Markov chains (*chaînes simples*) were considered. Bernstein's (1926) theorem (§9, Fundamental Lemma) can be considered a first central limit theorem for (discrete-time) semimartingales, deriving asymptotic normality from convergence of the predictable characteristics of the process. Lévy (1935) also studied martingales as part of his path-breaking work on central limit theory, and again proved a central limit theorem for discrete-time martingales (still under the name *variables enchaînées*).

The *word* martingale has a much longer prehistory. Leaving early equestrian, nautical, and Rabelaisian uses aside (see Rabelais, 1542), it was first used in an applied probability context to denote a betting system at Monte Carlo in which one counts on a long run of one color being very likely to be compensated soon by a run of the other color. The essence of the mathematical concept of martingale is, however, that this does not work: In a fair game, whatever betting system one uses (dependent on past observation), the fact is not altered that your expected winnings remain zero; in other words, integration of a predictable process with respect to a martingale produces a martingale. A version of this result was already given by Bernstein (1926, §11, Example).

The same idea was the basis of Richard von Mises' notion of a collective as a foundation for probability. A collective is a sequence such that whatever predictable rule is used to select subsequences, the law of large numbers remains valid for them. The word martingale was introduced to mathematics by Ville (1939) in part of a research program to discredit the theory of collectives. Using martingale theory, Ville showed that collectives exist for which the relative frequency of (say) ones in initial segments of the sequence *always* exceeds its limiting value; in other words, the object that was supposed to form the basis of probability theory did not itself satisfy the theorems of probability. This was just what the mathematical community had been waiting for, and the von Mises theory was abandoned in favor of the Kolmogorov axiomatization, though neither Kolmogorov nor von Mises saw the relevance of the example and something like collectives reappeared later in Kolmogorov's complexity theory. [See van Lambalgen (1987a,b) for an historical and foundational analysis. Kolmogorov and von Mises were interested in understanding randomness, not describing probability.]

## Section II.1

Our heuristic description of the theory has developed from Gill (1984) and Andersen and Borgan (1985).

## Sections II.2 and II.3

The pioneer of martingale theory, both in discrete and continuous time, was Doob (1940, 1953), who also gave the fundamental optional stopping theorem and the almost sure convergence theorem "for processes having the E-property" as they were first called.

Stochastic integration was pioneered by Itô (1944) [following the earlier work by Wiener (1923)] who gave a meaning to $\int_0^\tau H\,dW$ when $W$ is a Wiener process (or Brownian motion) and $H$ a random function; because the paths of $W$ are extremely erratic, this integral cannot be defined in a pathwise sense (unless $H$ is so nice that integration by parts can be used). The construction of $\int H\,dW$ used some basic properties of $W$ connected to what later became known as its predictable variation process: Not only is $W$ a martingale, but so also is $W^2 - \iota$. This implies that $\langle W \rangle = \iota$, or the compensator of the submartingale $W^2$ is the identity function $\iota$.

This observation, already made by Doob, motivated Meyer and his coworkers (Dellacherie, Doléans-Dade, and others) in Strasbourg in their search for the right conditions for what became known as the Doob–Meyer decomposition (Meyer, 1962), which they subsequently used to build a theory of stochastic integration. The story is long and complex and our remarks here obviously cannot do it justice.

Dellacherie's (1972) section theorem dealt with the highly delicate and fundamental measurability issues. Kunita and Watanabe (1967) furnished the idea of predictable covariation. The theory of stochastic integration was laid out by Meyer (1967) and put into its final form by Meyer (1976) [complete with the notions of *local martingale* going back to Itô and Watanabe (1965) and *semimartingale*, introduced by Doléans-Dade and Meyer (1970)].

Other celebrated results are Itô's (1951) formula, the theory of Doléans-Dade's (1970) exponential semimartingale, the characterization due independently to Dellacherie (1980) and to Bichteler (1979) of semimartingales as the largest class of processes for which a sensible theory of stochastic integration is possible, the so-called Burkholder–Davis–Gundy inequality, and so on. We do not make use of these results in this volume.

Theories of stochastic integration have also been developed from slightly different points of view by Métivier and Pellaumail (1980) among others.

Nowadays, the standard and complete work on continuous-time stochastic process theory, martingale theory, and stochastic integration is the book by Dellacherie and Meyer (1982), the second in a series of books covering many other topics at great depth. The theory was developed through alternative routes by Protter (1990) and by von Weizsäcker and Winkler (1990). Useful summaries were given by Jacod and Shiryayev (1987) and Daley and Vere-Jones (1988). In particular, for each topic, the book by Jacod and Shiryayev (1987) explicitly shows how the results simplify on specializing to the discrete-time case. The recent book by Liptser and Shiryayev (1989) also gives an excellent more extensive treatment of the theory, including central limit theory.

The hardest parts of the mathematical theory are contained in the Doob–Meyer decomposition and in Dellacherie's section theorems. If one only needs to apply the theory to a restricted class of applications, it may be possible to avoid some of these aspects. Jacobsen (1982), Fleming and Harrington (1991), and others each succeed in giving more elementary derivations of parts of the theory. Protter (1990), by starting from the Dellacherie–Bichteler characterization of semimartingale, was able to give most of the theory without recourse to the very heavy apparatus usually involved at the start of the "standard approach." Brown (1988), on the other hand, gave a complete proof of the Doob–Meyer decomposition using only more elementary tools. This proof was based on the idea of discretization and passage to the continuous limit (in discrete time, the Doob-Meyer decomposition is trivial).

We have made some minor adjustments to the usual definitions to suit our applications. As was explained in detail by Jacod (1979) or more briefly by Jacod and Shiryayev (1987, section I.1, especially Lemma I.1.19), it is not necessary to make the usual assumption of completeness of the filtration, which would otherwise be troublesome for statistical applications in which a whole family of probability measures is involved. Von Weizsäcker and Winkler (1990), in fact, showed how the whole theory can be set up without this assumption.

Our definition of localization is chosen to give the right meaning, both when the time interval $\mathcal{T}$ is open on the right and when it is closed on the right. Some other definitions also deviate from the usual ones to take care of the same possibilities. In our applications, martingales are nearly always local square integrable and of bounded variation (integrals of predictable processes with respect to counting process martingales), and in our survey of the theory, we concentrate on this case. Semimartingales are hardly needed and we have therefore neglected them, though in the modern theory they form the most fundamental object. The expert will also miss predictable stopping times from our survey.

## Section II.4

The idea of studying counting processes (or point processes) through their stochastic intensity was taken up seriously by Snyder (1972, 1975) with a view toward engineering applications in communication theory: filtering, prediction, smoothing, and so on (a different kind of filtering than the sort we study in Sections II.4.4 and III.4). A breakthrough came with the Berkeley thesis of Brémaud (1972) who realized that the previously rather heuristic notion of intensity could be made completely rigorous and connected to very powerful mathematical tools through noting that its operative value lies in the fact that *the integrated intensity process of the counting process coincides with its compensator*. The idea was taken up and further extended by Dolivo (1974), Boel, Varaiya, and Wong (1975a,b), and van Schuppen and Wong (1974)

among others. About the same time Jacod, spending a year in Princeton, came across Brémaud's thesis and wrote his (1973, 1975) papers connecting compensators and likelihood functions. The statistical implications of the theory was first realized by Aalen whose (1975) Berkeley thesis is the inspiration of this work; see Sections I.1 and I.2. The monograph of Jacobsen (1982) represents the first appearance of such material in book form. The book by Fleming and Harrington (1991) is the first extensive treatment aimed at applications in survival analysis. For surveys of counting process theory, see, for instance, Brémaud and Jacod (1977), Brémaud (1981), Karr (1986), and Daley and Vere-Jones (1988).

The idea of filtering a counting process in the sense used here is due to Aalen (1975). The notion of starting a counting process was introduced by Andersen et al. (1988) with help from M. Jacobsen. Fundamental results on the structure of point process filtrations were given by Courrège and Priouret (1965).

## Section II.5

As we indicated above, the history of martingales and the history of the central limit theorem are closely intertwined. After Bernstein's (1926) and Lévy's (1935) early work, the martingale central limit theorem for discrete-time martingales was taken up by Billingsley (1961), Brown (1971), Dvoretsky (1972), and McLeish (1974), among many others. Aalen's (1975) thesis [see also Aalen (1977)] contained a continuous-time martingale central limit theorem for counting process martingale integrals built on McLeish's theorem. This line of work was continued by Helland (1982, 1983) and Fleming and Harrington (1991). Rebolledo (1979) gave the first really general continuous-time martingale central limit theorem; see also Rebolledo (1980a). One of the key tools in his approach was Lenglart's (1977) inequality, which we also use time and time again. Rebolledo (1979) used his own tightness criterion, but a more powerful one was given by Aldous (1978a).

Independently of Rebolledo, authors such as Jacod and Memin (1980), Kłopotowski (1980), and Liptser and Shiryayev (1980), worked on the central limit problem for semimartingales, with the more general ultimate aim to develop theorems with as limiting process *any* semimartingale (not just Gaussian limits) and to develop *necessary* as well as sufficient conditions [for this latter, see also Rebolledo (1980b)]. The developments were surveyed by Shiryayev (1981). This line of work culminated (for the time being, at least), in the book by Jacod and Shiryayev (1987).

An interesting contribution was made by Aldous (1978b), unfortunately never published, who gave a weak convergence theory in which filtration-related properties converge too (e.g., predictable processes converge to predictable processes, martingales to martingales, and so on).

We present the version of the martingale central limit theorem relevant for our applications: the case of a vector of local square integrable martingales.

Time transformations for counting processes were treated by Aalen (1975) and Aalen and Hoem (1978). Independently, Kurtz (1983) used the idea of time transformation of a counting process to a Poisson process to obtain very powerful laws of large numbers and central limit theorems. The conditions for convergence in probability of integrals were proposed by Helland (1983) and Gill (1983b), at the same meeting.

## Section II.6

Product-integration was introduced by Volterra (1887) as part of the theory of differential equations. Though its significance in probability and statistics was seen by such authors as Arley (1943) and Dobrushin (1953), it somehow never quite became a well-known theory, and it has been reinvented many times. Cox (1972) gives a side reference to product-integration (which he was aware of through Arley's contribution).

Johansen (1977, reprinted 1986) gave a treatment of product-integration aimed at applications in Markov processes. The monograph on product-integration by Dollard and Friedman (1979), though a mine of information on many aspects of the theory, neglects the continuous–discrete interplay which is so important for our applications. The presentation here is based on Gill and Johansen (1990) who gave complete proofs and many further results. Our notation for product-integral (a script capital letter pi) is due to them.

Readers familiar with stochastic analysis will recognize the product-integral as Doléans–Dade's exponential semimartingale: the solution $Y = \mathscr{E}(X)$ of the stochastic differential equation $\mathrm{d}Y(t) = Y(t-)\,\mathrm{d}X(t)$, with initial condition $Y(0) = 1$.

## Section II.7

Likelihood representations for general counting process models were first given by Jacod (1973, 1975). Use of product-integral notation to make the otherwise rather involved formulas more transparent goes back to Johansen (1983). Partial likelihood was invented by Cox (1975) to retrospectively justify the statistical methods he had proposed (Cox, 1972) for analyzing his regression model for censored survival times; see Section VII.2. Kalbfleisch and Prentice (1980) used the idea to informally discuss likelihood constructions for survival data. Our discussion builds on the work of Arjas and Haara (1984) who were the first to formulate the notions of noninformative and independent censoring rigorously, by formulating these notions in terms of likelihoods and intensity processes of general marked point processes. Arjas (1989) and Arjas, Haara, and Norros (1992) developed the ideas further.

Asymptotic statistical theory for partial likelihood was developed by Wong (1986). Slud (1991) derived our "continuous" partial likelihoods rigor-

ously as limits of partial likelihood based on factorizations of an experiment into a finite number of terms. Jacod (1987, 1990a,b), inspired by an early manuscript version of the present material, shows how asymptotic properties of partial likelihoods to some extent have the same striking consequences as those of full likelihood (contiguity and local asymptotic normality; see Chapter VIII).

## Section II.8

The material on the functional delta-method given here is largely taken from the paper by Gill (1989), whose own work was inspired by Reeds' (1976) Harvard thesis. The word delta-method is due to C.R. Rao. The ideas of the functional delta-method go back to von Mises (1947), and many other authors built on his work. Reeds (1976) was the first to point out that an elegant von Mises-type theory could be based on the notions of Hadamard or compact differentiability, rather than the Gâteaux- (directional differentiation) or Fréchet- (bounded differentiability) based theories favored by other authors. Incidentally, both Fréchet and Gâteaux were pupils of Hadamard (another famous pupil being P. Lévy who also made some contributions to this area). The idea of compact differentiation can be found in the early work of Hadamard, but it was Fréchet (1937) who recognized its significance and named it after his teacher, after first having developed his own notion of differentiation.

Reeds' (1976) work, not easily available, has been largely neglected or criticized, partly because of his rather polemic style. His main result, a compact differentiability-based proof of the asymptotic normality of M-estimators, was corrected by Heesterman and Gill (1992). Further developments were given by Sheehy and Wellner (1990a,b) using an even further generalized weak convergence theory (van der Vaart and Wellner, 1990, Pollard, 1990).

## Section II.9

Historical and bibliographic surveys of stochastic analysis can be found in Dellacherie and Meyer (1982), Liptser and Shiryayev (1989), and Protter (1990). For counting processes, see Brémaud (1981); for product-integration, see Gill and Johansen (1990); for the functional delta-method, see Reeds (1976).