Inference for Stochastic Processes

Robert L. Wolpert*

Revised: June 19, 2005

Introduction

A **stochastic process** is a family $\{X_t\}$ of real-valued random variables, all defined on the same probability space $(\Omega, \mathcal{F}, \mathsf{P})$ so that it will make sense to talk about their joint distribution. In many applications the set $\mathcal{T}$ of indices t is infinite (common examples include the non-negative integers $\mathbb{Z}_+$ or reals $\mathbb{R}_+$), so the random variables $\{X_t\}$ won't have a joint probability density function or probability mass function– we must find some other way to specify the joint distribution, and to make inference about it. In this class we will do this for four specific classes of stochastic processes:

1. Markov Chains,
2. Diffusions,
3. Lévy Processes, and
4. Gaussian Processes.

Three of these classes are examples of *Markov Processes*, so it is worth while introducing a few tools useful in the study of all of them. First, an example.

*Robert L. Wolpert is Professor of Statistics and Decision Sciences and Professor of the Environment at Duke University, Durham, NC, 27708-0251.

Example: Bernoulli Sequence

Let $\Omega = (0, 1]$ be the unit interval and, for $\omega \in \Omega$ and $n \in \mathbb{N} = \{1, 2, ...\}$, let $\beta_n(\omega)$ be the n^{th} bit in the binary expansion of ω,

$$\beta_n(\omega) = \lfloor 2^n \omega \rfloor \pmod 1$$

If $\mathsf{P}(d\omega)$ is Lebesgue measure, then $\{\beta_n\}$ are independent Bernoulli random variables, each equal to one with probability $\frac{1}{2}$ and otherwise zero. We may use these to construct many interesting examples. For any $0 \leq p \leq 1$ it is possible to construct a measure $\mathsf{P}_p(d\omega)$ on Ω for which the $\{\beta_n\}$ are independent and identically distributed with $\mathsf{P}_p[\beta_n = 1] = p$.

A Little Probability Theory

Probability and Statistics are complementary sciences for studying the world of uncertain phenomena. In *Probability*, we (pretend that we) know all about the mechanism governing some unpredictable observations, and we compute the probabilities of events we have not yet seen; in *Statistics* we (pretend that we) know which events have occured and which have not, and we try to make inference about what must have been the mechanism governing those observations. In the Bernoulli setting, for example, probability theory allows us to compute that the probability of an even number of 0's before the first 1 is 2/3, if the bits are zero or one with probability one-half each, while the probability that the total number of ones among the first 1000 bits is 500 ± 20 is $\mathsf{P}[480 \leq X \leq 520 \mid p = 0.5] = 0.80534$, while statistics allows us to infer, upon seeing 600 ones among the first thousand bits, that it is unlikely that the $\{\beta_n\}$ are independent with $\mathsf{P}[\beta_n = 1] = \frac{1}{2}$. More precisely, we can infer from a Bayesian perspective that IF the β_n are *i.i.d.* Bernoulli with some probability parameter p, that $\mathsf{P}[p \in (0.574, 0.625) \mid X = 600] \approx 0.90$ and $\mathsf{P}[p \leq 0.5 \mid X = 600] \approx 1.1 \times 10^{-10}$, while from a frequentist perspective we compute that $\mathsf{P}[X \geq 600 \mid p = \frac{1}{2}] = 1.36 \times 10^{-10}$ and infer that it would be a miracle to see so many ones if $p = \frac{1}{2}$.

Probability Theory concerns what might happen in a "random experiment." It is chiefly concerned with *events*, which we think of informally as "things that might happen, and then again might not," and with *random variables*, or "numbers that depend on chance." More formally we represent everything that might happen in the experiment by the elements of some

set Ω called the *sample space*, and represent events by subsets $E \subset \Omega$ and random variables by real-valued functions $X : \Omega \to \mathbb{R}$, with the notion that "performing the experiment" is the same as choosing some element $\omega \in \Omega$, whereupon an event E has occurred if $\omega \in E$ and otherwise has not, and that the value of a random variable is observed to be $X(\omega)$.

Let $\mathcal{F}$ denote the collection of all the Events E whose probability $\mathsf{P}[E]$ we can compute. In some cases this might be all subsets $E \subset \Omega$, but perhaps that is too ambitious. Certainly we can compute $\mathsf{P}[\Omega] = 1$ (the probability that *anything at all* happens has to be one), and we can compute $\mathsf{P}[E^c]$ if we can compute $\mathsf{P}[E]$ (it's just $\mathsf{P}[E^c] = 1 - \mathsf{P}[E]$, since necessarily either E occurs or it doesn't), and we will want to be able to compute $\mathsf{P}[E_1 \cap E_2]$ and $\mathsf{P}[E_1 \cup E_2]$ for events E_1 and E_2. Any collection of sets that satisfies these three rules is called an *algebra*. If it satisfies the somewhat more stringent rule that it contains $\cup E_i$ (the event that at least one of the $\{E_i\}$ occurs) and $\cap E_i$ (the event that all of the $\{E_i\}$ occur) for any countable collection $\{E_i\} \subset \mathcal{F}$, then $\mathcal{F}$ is called a σ-algebra.

We wish to assign a *probability* $\mathsf{P}(E)$ to each event $E \in \mathcal{F}$ in ways that satisfy the "obvious" rules

1. $\mathsf{P}[E] \geq 0$ for every $E \in \mathcal{F}$;
2. $\mathsf{P}[\Omega] = 1$;
3. $\mathsf{P}[E_1 \cup E_2] = \mathsf{P}[E_1] + \mathsf{P}[E_2]$, if $E_1 \cap E_2 = \emptyset$. It turns out to be useful to strengthen this to countable unions, *i.e.*,
   $\mathsf{P}[\cup E_i] = \sum \mathsf{P}[E_i]$, if $E_i \cap E_j = \emptyset \, \forall i \neq j$.

A **probability** is a real-valued function $\mathsf{P} : \mathcal{F} \to \mathbb{R}$ that satisfies these three rules.

The **probability distribution** of a random variable X is the probability assignment $\mu_X(B) = \mathsf{P}[X \in B]$ for sets $B \subset \mathbb{R}$; for this to even make sense we will need $\{\omega : X(\omega) \in B\} = X^{-1}(B)$ to be an event. It turns out that the collection of sets $B \subset \mathbb{R}$ for which $X^{-1}(B) \in \mathcal{F}$ is a σ-algebra, so requiring this for intervals $B = (a, b]$ (which we would need just to define the CDF $F_X(b) = \mathsf{P}[X \leq b]$) is just the same as asking that $X^{-1}(\mathcal{B}) \subset \mathcal{F}$ for the entire collection $\mathcal{B}$ of "Borel sets," the smallest σ-algebra of sets $B \subset \mathbb{R}$ that includes all the open sets (or, equivalently, all the intervals).

For any collection $\{E_\alpha\}$ of subsets $E_\alpha \subset \Omega$ there is a smallest σ-algebra $\sigma\big(\{E_\alpha\}\big)$ that contains them all; for example, the "Borel Sets" $\mathcal{B}$ of any

topological space are defined to be the smallest σ-algebra containing all the open sets. Similarly for any collection of random variables $\{X_\alpha\}$ there is a smallest σ-algebra $\mathcal{G} = \sigma\big(\{X_\alpha\}\big) \subset \mathcal{F}$ that contains every ${X_\alpha}^{-1}(\mathcal{B})$.

If Y is any L^1 random variable (*i.e.*, one such that $\mathsf{E}[|Y|] < \infty$) then we define the *conditional expectation of Y, given $\mathcal{G}$*, denoted $\mathsf{E}[Y \mid \mathcal{G}]$, to be the best approximation to Y from among all possible functions of $\{X_\alpha\}$. In particular, if Y is already one of the $\{X_\alpha\}$, or some function of finitely many $\{X_{\alpha_j}\}$, or a limit of such things, then $E[Y \mid \mathcal{G}]$ is just Y itself. More generally, this conditional expectation is defined to be any random variable satisfying the two conditions

$$\begin{aligned} \mathsf{E}[Y \mid \mathcal{G}] &\in \mathcal{G} \qquad \text{(short-hand for } (\mathsf{E}[Y \mid \mathcal{G}])^{-1}(\mathcal{B}) \subset \mathcal{G}) \\ \mathsf{E}[(\mathsf{E}[Y \mid \mathcal{G}] - Y)1_G] &= 0 \qquad \text{for all sets } G \in \mathcal{G} \end{aligned}$$

The first condition says that $\mathsf{E}[Y \mid \mathcal{G}]$ has to be a limit of functions of finitely many of the $\{X_\alpha\}$, while the second condition says that the approximation is so good that the average error is zero over any event in $\mathcal{G}$ or, equivalently, that the error is orthogonal to every bounded function of the $\{X_\alpha\}$.

This reduces to the familiar notion of conditional expectation

$$E[Y|X_1, ..., X_m] = \frac{\int y\, f(\vec{x}, y)\, dy}{\int f(\vec{x}, y)\, dy}$$

when $\vec{X} = (X_1, ..., X_m)$ is finite dimensional with a joint density function, so $\mathcal{G} = \sigma(X_1, ..., X_m)$ depends on only finitely many random variables, but extends the idea to conditioning on infinitely many random variables (even uncountably many). Similarly it extends the notions of conditional probability $\mathsf{P}[A|B]$, with $\mathcal{G} = \sigma(1_B) = \{\emptyset, B, B^c, \Omega\}$ and

$$\mathsf{P}[A|\mathcal{G}](\omega) = \begin{cases} \mathsf{P}[A \mid B\ ] = \mathsf{P}[A \cap B\ ]/\mathsf{P}[B\ ] & \text{for } \omega \in B, \\ \mathsf{P}[A \mid B^c] = \mathsf{P}[A \cap B^c]/\mathsf{P}[B^c] & \text{for } \omega \notin B. \end{cases}$$

In many applications there will be an obvious candidate for this optimal predictor; an example is the case of martingales.

Martingales

In cases where $\mathcal{T} \subset \mathbb{R}$ is one-dimensional the indices are naturally ordered. We often speak of $t \in \mathcal{T}$ as "time" and think of the "past" at any time $t \in \mathcal{T}$

as including whatever we can observe at times $s \le t$. Let $\mathcal{F}_t^X$ represent the "past" at time $t \in \mathcal{T}$ of the process X_t; formally, it consists of all events $E \in \mathcal{F}$ that depend only on $X_s : s \le t$, and in particular contains all the events ${X_t}^{-1}(\mathcal{B})$ that depend just on the process at time t. More generally a **filtration** is *any* increasing family $\mathcal{F}_t \subset \mathcal{F}$ of σ-algebras, and X is said to be *adapted* to $\mathcal{F}_t$ if ${X_t}^{-1}(\mathcal{B}) \subset \mathcal{F}_t$ for every $t \in \mathcal{T}$. Then $\mathcal{F}_t$ will contain the past of X_t but perhaps will also contain the past of other processes as well. In the Bernoulli example,

$$\mathcal{F}_t = \left\{\text{all unions of intervals of the form } (i/2^t, j/2^t], \qquad 0 \le i < j \le 2^t\right\}$$

is just the collection of all unions of half-open intervals with dyadic rational endpoints of order t or less, and the conditional expectation $\mathsf{E}[Y \mid \mathcal{F}_t]$ of any integrable function $Y \in L^1(0,1)$ is just the piecewise-constant simple function whose constant value on each dyadic interval of order t is given by Y's average value over that interval,

$$\mathsf{E}[Y \mid \mathcal{F}_t](\omega) = 2^{-t} \int_{i/2^t}^{(i+1)/2^t} Y(\omega)\, d\omega, \qquad \frac{i}{2^t} < \omega \le \frac{i+1}{2^t}$$

Let $\mathcal{F}_t$ be a filtration. A real-valued stochastic process M_t is said to be a **martingale** for $\mathcal{F}_t$ if

- $\mathsf{E}[|M_t|] < \infty$ for every $t \in \mathcal{T}$
- $\mathsf{E}[M_t \mid \mathcal{F}_s] = M_s$ for every $s \le t \in \mathcal{T}$.

The second condition asserts that the best predictor of M_t available, from among all possible functions of $\{M_r : r \le s\}$ or of any other random variables Y with $Y^{-1}(\mathcal{B}) \subset \mathcal{F}_s$, is M_s itself. This says in a very strong way that M_t is "conditionally constant", that on average it neither increases nor decreases over time.

It will follow that $\mathsf{E}[M_t] = \mathsf{E}[M_0]$, of course, but something much stronger is true; $\mathsf{E}[M_\tau] = \mathsf{E}[M_0]$ even at *random times* $\tau \in \mathcal{T}$, so long as the random time "depends only on the past." More formally, $\tau : \Omega \to \mathcal{T}$ is called a *Stopping Time* (or sometimes a *Markov time*) if

$$\{\omega : \tau(\omega) \le t\} \in \mathcal{F}_t$$

for every $t \in \mathcal{T}$— *i.e.*, that the event "$[\tau \le t]$" depends only on the $\{X_s\}$ for $s \le t$. A stopping time in a gambling game would be a rule for when

to quit that depends only on the past and present; this forbids quitting "just before we lose" (too bad!). Martingales have at least three remarkable properties that make them ideal tools for studying stochastic processes and their inference:

- **Doob's Optional Stopping Theorem:**
  If M_t is a martingale and τ a stopping time then $M_{t\wedge\tau}$ is a martingale, too. This implies that $\mathsf{E}[M_\tau] = \mathsf{E}[M_0]$ for any stopping time τ, if either $\mathsf{E}[\tau] < \infty$ or if $\{M_t\}$ is UI. (Note: This could be taken as an alternate *definition* of Martingale).

- **Martingale Convergence Theorem:**
  If M_t is Uniformly Integrable then there exists an L^1 random variable Z s.t. $M_t \to Z$ *a.s.* as $t \to \infty$, and $M_t = E[Z|\mathcal{F}_t]$ for all $t < \infty$. One sufficient condition for uniform integrability is that $|M_t| \leq Y$ for some $Y \in L^1$; another is that $\mathsf{E}[|M_t|^p] \leq K$ for some $p > 1$ and $K < \infty$.

- **Martingale Maximal Inequality:**
  For any $0 \leq t \leq \infty$ set $M_t^* \equiv \sup[M_s : 0 \leq s \leq t]$, the maximum value of M_s on the interval $s \leq t$. Let $p > 1$ and set $q = p/(p-1)$ (so $1/p + 1/q = 1$). Set $\|X\|_p \equiv (\mathsf{E}|X|^p)^{1/p}$ (the ordinary L^p norm). Then for any $c > 0$,

$$\begin{aligned} \mathsf{P}[|M_t^*| > c] &\leq E[|M_t|]/c \\ \|M_t^*\|_p &\leq q \sup_{s\leq t} \|M_s\|_p \end{aligned}$$

Notice that the Maximal Inequality is reminiscent of the Markov Inequality

$$\mathsf{P}[|Y| > c] \leq E[|Y|]/c$$

true for any L^1 random variable, but it is much stronger— the bound is not only on $|M_t|$, but on the maximum $|M_t^*|$.

Examples

Recall the Bernoulli random variables $\beta_n \sim \mathsf{Bi}(1, \frac{1}{2})$ constructed above. Define a symmetric random walk starting at $x \in \mathbb{Z}$ by

$$S_t \equiv x + \sum_{n=1}^{t} (2\beta_n - 1),$$

and let $\mathcal{F}_t = \sigma(S_n : n \le t) = \sigma(\beta_n : n \le t)$. Evidently S_t begins at $S_0 = x$ and at each time t takes a unit step up if $\beta_t = 1$ and takes a unit step down if $\beta_t = 0$, equally likely events. It follows that S_t is a martingale, of course, but so too is $M(t) = (S_t)^2 - t$ (check this).

For $0 < p < 1$ under the probability assignment P_p that makes the β_n into *i.i.d.* $\mathsf{Bi}(1,p)$ random variables, S_t will become a biased random walk that steps to the right (*i.e.*, up) with probability p and to the left with probability $q = 1 - p$. If $p \neq \frac{1}{2}$ then S_t is no longer a martingale, but both $M_0(t) = (q/p)^{S_t}$ and $M_1 = S_t - (p - q)t$ are martingales.

These lead to a simple and elegant solution of the famous *Gambler's Ruin* problem: for $a \le x \le b$ let τ be the first time that S_t leaves the open interval (a, b), $\tau \equiv \min\{t \in \mathcal{T} : S_t \neq (a, b)\}$, *i.e.*, the first hitting time of $\{a, b\}$, and let $f(x) = \mathsf{P}[S_\tau = b]$ be the probability that the interval is exited to the right. This function $f(x)$ may be calculated using Doob's Optional Sampling Theorem.

For the symmetric ($p = \frac{1}{2}$) case,

$$\begin{aligned} \mathsf{E}[S_\tau \mid S_0 = x] &= [f(x)]b + [1 - f(x)]a \\ &= a + f(x)(b - a) \\ &= x \qquad \text{(by Doob's O.S.T.), so} \\ f(x)(b - a) &= x - a \text{ and } f(x) = \frac{x - a}{b - a}. \end{aligned}$$

For example, a gambler with a \$100 fortune has a probability of $f(100) = 100/110$ of winning \$10 before going broke, when playing a fair game betting \$1 each play with even odds (here $a = 0$, $x = 100$, and $b = 110$). The expected time to complete play is available too:

$$\begin{aligned} \mathsf{E}[(S_\tau)^2 - \tau \mid S_0 = x] &= [f(x)]b^2 + [1 - f(x)]a^2 - \mathsf{E}[\tau \mid S_0 = x] \\ &= a^2 + f(x)(b^2 - a^2) - \mathsf{E}[\tau \mid S_0 = x] \\ &= x^2 \qquad \text{(by Doob's O.S.T.), so} \\ \mathsf{E}[\tau \mid S_0 = x] &= \frac{(x - a)(b^2 - a^2)}{b - a} - (x^2 - a^2) = (b - x)(x - a), \end{aligned}$$

or 1000 plays in the gambling example above.

For the asymmetric case,

$$\begin{aligned}
\mathsf{E}[(q/p)^{S_\tau} \mid S_0 = x] &= [f(x)](q/p)^b + [1 - f(x)](q/p)^a \\
&= (q/p)^a + f(x)[(q/p)^b - (q/p)^a] \\
&= (q/p)^x \text{ by Doob's O.S.T., so} \\
f(x) &= \frac{(q/p)^x - (q/p)^a}{(q/p)^b - (q/p)^a} \\
&= \frac{(p/q)^{b-x} - (p/q)^{b-a}}{1 - (p/q)^{b-a}} \\
&\approx (p/q)^{b-x} \text{ if } (p/q)^{b-a} \approx 0
\end{aligned}$$

Betting on red or black roulette in the U.S. has probability $p = 9/19$ of winning and $q = 10/19$ of losing, so a gambler with a \$100 fortune has a probability of

$$f(100) = \frac{(9/10)^{10} - (9/10)^{110}}{1 - (9/10)^{110}} \approx (9/10)^{10} = .3487$$

of winning \$10 before going broke, when betting \$1 on red or black at U.S. roulette. Actually he or she has the same odds (to four decimal places) of winning \$10 before going broke beginning with the largest fortune of anyone on earth (presently about \$50B)! The expected time to complete play is available once again:

$$\begin{aligned}
\mathsf{E}[S_\tau - (p-q)\tau \mid S_0 = x] &= [f(x)]b + [1 - f(x)]a - (p-q)\mathsf{E}[\tau \mid S_0 = x] \\
&= x \text{ by Doob's O.S.T., so} \\
\mathsf{E}[\tau \mid S_0 = x] &= \frac{[f(x)]b + [1 - f(x)]a - x}{p - q} \\
&= \frac{(b-x)[(p/q)^{b-x} - (p/q)^{b-a}] - (x-a)[1 - (p/q)^{b-x}]}{(p-q)[1 - (p/q)^{b-a}]} \\
&\approx \frac{(x-a) - (b-a)(p/q)^{b-x}}{q-p} \text{ if } (p/q)^{b-a} \approx 0
\end{aligned}$$

or about $\mathsf{E}[\tau \mid S_0 = 100] = [100 - 110(9/10)^{10}]/[1/19] \approx 1171$ for our ill-fated gambler.

The problem of doubling our money is even more striking: Beginning with a stake of $x = \$50$, the chance of doubling it to reach $b = 100$ before

losing it all (so $a = 0$) is $f(50) = \frac{1}{2}$ in the fair game, with an expectation of taking $50^2 = 2500$ plays to complete the game, while the house edge afforded by the (green) "0" and "00" possibilities at U.S. roulette drop the gambler's chances to $f(50) \approx (0.9)^{50} = .00515$, with an expected playing time of only 940 plays. Note that S_t is a random walk whose steps have expectation $\mathsf{E}[S_t - S_{t-1}] = \mathsf{E}[2\beta_t - 1] = (p - q) = -1/19$, so on average it would take about $50 \times 19 = 950$ steps to fall from 50 to zero; the expected length of the game is a little shorter than that, due to the approximately one in two-hundred chance of winning.

Now we are ready to begin studying inference for Markov Processes; we begin by introducing Markov Chains.