MAST30001 Stochastic Modelling

Lecturer: Nathan Ross

Administration

- LMS - announcements, grades, course documents
- Lectures/Practicals
- Student-staff liaison committee (SSLC) representative

Modelling

We develop an *imitation* of the system. It could be, for example,

- a small replica of a marina development,
- a set of equations describing the relations between stock prices,
- a computer simulation that reproduces a complex system (think: the paths of planets in the solar system).

We use a model

- to understand the evolution of a system,
- to understand how ‘outputs’ relate to ‘inputs’, and
- to decide how to influence a system.

Why do we model?

We want to understand how a complex system works. Real-world experimentation can be

- too slow,
- too expensive,
- possibly too dangerous,
- may not deliver insight.

The alternative is to build a physical, mathematical or computational model that captures the essence of the system that we are interested in (think: NASA).

Why a stochastic model?

We want to model such things as

- traffic in the Internet
- stock prices and their derivatives
- waiting times in healthcare queues
- reliability of multicomponent systems
- interacting populations
- epidemics

where the effects of randomness cannot be ignored.

Good mathematical models

- capture the non-trivial behaviour of a system,
- are as simple as possible,
- replicate empirical observations,
- are tractable - they can be analysed to derive the quantities of interest, and
- can be used to help make decisions.

Stochastic modelling

Stochastic modelling is about the study of random experiments. For example,

- toss a coin once, toss a coin twice, toss a coin infinitely-many times
- the lifetime of a randomly selected battery (quality control)
- the operation of a queue over the time interval $[0, \infty)$ (service)
- the changes in the US dollar - Australian dollar exchange rate from 2006 onwards (finance)
- the positions of all iphones that make connections to a particular telecommunications company over the course of one hour (wireless tower placement)
- the network "friend" structure of Facebook (ad revenue)

Stochastic modelling

We study a random experiment in the context of a Probability Space $(\Omega, \mathcal{F}, P)$. Here,

- the sample space Ω is the set of all possible outcomes of our random experiment,
- the class of events $\mathcal{F}$ is a set of subsets of Ω. We view these as events we can *see* or *measure*, and
- P is a probability measure defined on the elements of $\mathcal{F}$.

The sample space Ω

We need to think about the sets of possible outcomes for the random experiments. For those discussed above, these could be

- $\{H, T\}$, $\{(H, H), (H, T), (T, H), (T, T)\}$, the set of all infinite sequences of Hs and Ts.
- $[0, \infty)$.
- the set of piecewise-constant functions from $[0, \infty)$ to $\mathcal{Z}_+$.
- the set of continuous functions from $[0, \infty)$ to $\mathbb{R}_+$.
- $\bigcup_{n=0}^{\infty}\{(x_1, y_1) \ldots (x_n, y_n)\}$, giving locations of the phones when they connected.
- Set of simple networks with number of vertices equal to the number of users: edges connect friends.

Review of basic notions of set theory

- $A \subset B$.
  - A is a subset of B or if A occurs, then B occurs.
- $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\} = B \cup A$.
  - Union of sets (events): at least one occurs.
  - $A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i$.
- $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\} = B \cap A = AB$.
  - Intersection of sets (events): both occur.
  - $A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i$.
- $A^c = \{\omega \in \Omega : w \notin A\}$
  - Complement of a set/event: event doesn't occur.
- $\emptyset$: the empty set or impossible event.

The class of events $\mathcal{F}$

- For discrete sample spaces, $\mathcal{F}$ is typically the set of all subsets.

  Example: Toss a coin once, $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$
- For continuous state spaces, the situation is more complicated:

The class of events $\mathcal{F}$

- S equals circle of radius 1.
- We say two points on S are in the same family if you can get from one to the other by taking steps of arclength 1 around the circle.
- Each family chooses a single member to be head.
- If X is a point chosen uniformly at random from the circle, what is the chance X is the head of its family?

The class of events $\mathcal{F}$

- $A = \{X \text{ is head of its family}\}$.
- $A_i = \{X \text{ is } i \text{ steps clockwise from its family head}\}$.
- $B_i = \{X \text{ is } i \text{ steps counterclockwise from its family head}\}$.
- By uniformity, $P(A) = P(A_i) = P(B_i)$, BUT
- law of total probability:

$$1 = P(A) + \sum_{i=1}^{\infty} (P(A_i) + P(B_i))!$$

The issue is that the event A is not one we can *see* or *measure* so should not be included in $\mathcal{F}$.

The class of events $\mathcal{F}$

These kinds of issues are technical to resolve and are dealt with in later probability or analysis subjects which use *measure theory*.

The probability measure P

The probability measure P on $(\Omega, \mathcal{F})$ is a set function from $\mathcal{F}$ satisfying

P1. $P(A) \geq 0$ for all $A \in \mathcal{F}$

[probabilities measure long run %'s or certainty]

P2. $P(\Omega) = 1$

[There is a 100% chance something happens]

P3. Countable additivity: if A_1, $A_2 \cdots$ are disjoint events in $\mathcal{F}$, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

[Think about it in terms of frequencies]

How do we specify P?

The modelling process consists of

- defining the values of $P(A)$ for some 'basic events' in $A \in \mathcal{F}$,
- deriving $P(B)$ for the other 'unknown' more complicated events in $B \in \mathcal{F}$ from the axioms above.

Example: Toss a fair coin 1000 times. Any length 1000 sequence of H's and T's has chance 2^{-1000}.

- What is the chance there are more than 600 H's in the sequence?
- What is the chance the first time the proportion of heads exceeds the proportion of tails occurs after toss 20?

Properties of P

- $P(\emptyset) = 0$.
- $P(A^c) = 1 - P(A)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional probability

Let $A, B \in \mathcal{F}$ be events with $P(B) > 0$. Supposing we know that B occurred, how likely is A given that information? That is, what is the conditional probability $P(A|B)$?

For a frequency interpretation, consider the situation where we have n trials and B has occurred n_B times. What is the relative frequency of A in these n_B trials? The answer is

$$\frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n} \sim \frac{P(A\cap B)}{P(B)}.$$

Hence, we define

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

We need a more sophisticated definition if we want to define the conditional probability $P(A|B)$ when $P(B) = 0$.

Example:

Tickets are drawn consecutively and *without replacement* from a box of tickets numbered $1 - 10$. What is the chance the second ticket is even numbered given the first is

- even?
- labelled 3?

Bayes' formula

Let $B_1, B_2, \cdots, B_n$ be mutually exclusive events with $A \subset \bigcup_{j=1}^{n} B_j$, then

$$P(A) = \sum_{j=1}^{n} P(A|B_j)P(B_j).$$

With the same assumptions as for the Law of Total Probability,

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}.$$

Example:

A disease affects $1/1000$ newborns and shortly after birth a baby is screened for this disease using a cheap test that has a 2% false positive rate (the test has no false negatives). If the baby tests positive, what is the chance it has the disease?

Independent events

Events A and B are said to be independent if

$$P(A \cap B) = P(A)P(B).$$

If $\mathbb{P}(B) \neq 0$ or $\mathbb{P}(A) \neq 0$ then this is the same as $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Events $A_1, \cdots, A_n$ are independent if, for any subset $\{i_1, ..., i_k\}$ of $\{1, ..., n\}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k}).$$

Random variables

A random variable (rv) on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$.

Usually, we want to talk about the probabilities that the values of random variables lie in sets of the form $(a, b) = \{x : a < x < b\}$. The smallest σ-algebra of subsets of $\mathbb{R}$ that contains these sets is called the set $\mathcal{B}(\mathbb{R})$ of Borel sets, after Emile Borel (1871-1956).

The probability that $X \in (a, b)$ is the probability of the subset $\{\omega : X(\omega) \in (a, b)\}$. In order for this to make sense, we need this set to be in $\mathcal{F}$ and we require this condition for all $a < b$ and we say the function X is *measurable* with respect to $\mathcal{F}$.

So X is measurable with respect to $\mathcal{F}$ if $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ for all Borel sets $B \subset \mathbb{R}$.

Distribution Functions

The function $F_X(t) = P(X \leq t) = P(\{\omega : X(\omega) \in (-\infty, t]\}$ that maps R to $[0, 1]$ is called the distribution function of the random variable X.

Any distribution function F

F1. is non-decreasing,

F2. is such that $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$, and

F3. is 'right-continuous', that is $\lim_{h \to 0^+} F(t + h) = F(t)$ for all t.

Distribution Functions

We say that

- the random variable X is discrete if it can take only countably-many values, with $P(X = x_i) = p_i > 0$ and $\sum_i p_i = 1$. Its distribution function $F_X(t)$ is commonly a step function.
- the random variable X is continuous if $F_X(t)$ is absolutely continuous, that is if there exists a function $f_X(t)$ that maps R to R_+ such that $F_X(t) = \int_{-\infty}^{t} f_X(u)du$.

A mixed random variable has some points that have positive probability and also some continuous parts.

Examples of distributions

- Examples of discrete random variables: binomial, Poisson, geometric, negative binomial, discrete uniform
  http://en.wikipedia.org/wiki/Category:Discrete_distributions
- Examples of continuous random variables: normal, exponential, gamma, beta, uniform on an interval (a, b)
  http://en.wikipedia.org/wiki/Category:Continuous_distributions

Random Vectors

A random vector $\mathbf{X} = (X_1, ..., X_d)$ is a measurable mapping of $(\Omega, \mathcal{F})$ to $\mathbb{R}^d$, that is, for each Borel set $B \subset \mathbb{R}^d$, $\{\omega : X(\omega) \in B\} \in \mathcal{F}$.

The distribution function of a random vector is

$$F_{\mathbf{X}}(\mathbf{t}) = P(X_1 \leq t_1, \cdots, X_d \leq t_d), \ \mathbf{t} = (t_1, \cdots, t_d) \in \mathbf{R}^d.$$

It follows that

$$\begin{aligned} & P(s_1 < X_1 \leq t_1, s_2 < X_2 \leq t_2) \\ = \ & F(t_1, t_2) - F(s_1, t_2) - F(t_1, s_2) + F(s_1, s_2). \end{aligned}$$

Independent Random Variables

The random variables $X_1, \cdots, X_d$ are called independent if $F_{\mathbf{X}}(\mathbf{t}) = F_{X_1}(t_1) \times \cdots \times F_{X_d}(t_d)$ for all $\mathbf{t} = (t_1, \cdots, t_d)$.
Equivalently,

- the events $\{X_1 \in B_1\}$, $\cdots$, $\{X_d \in B_d\}$ are independent for all Borel sets $B_1, \cdots, B_d \subset R$,
- or, in the absolutely-continuous case, $f_{\mathbf{X}}(\mathbf{t}) = f_{X_1}(t_1) \times \cdots \times f_{X_d}(t_d)$ for all $\mathbf{t} = (t_1, \cdots, t_d)$.

Revision Exercise

For bivariate random variables (X, Y) with density functions

- $f(x, y) = 2x + 2y - 4xy$ for $0 < x < 1$, $0 < y < 1$, and
- $f(x, y) = 4 - 4x - 4y + 8xy$ for $0 < x < 1$, $0 < y < 1$, $0 < x + y < 1$,
    - check f is a true density.
    - find the marginal probability density functions $f_X(x)$ and $f_Y(y)$,
    - find the probability density function of Y conditional on the value of X.

Expectation of X

For a discrete, continuous or mixed random variable X that takes on values in the set S_X, the expectation of X is

$$E(X) = \int_{S_X} x dF_X(x)$$

The integral on the right hand side is a Lebesgue-Stieltjes integral. It can be evaluated as

$$= \begin{cases} \sum_{i=1}^{\infty} x_i P(X = x_i), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx, & \text{if } X \text{ is absolutely continuous.} \end{cases}$$

In second year, we required that the integral be absolutely convergent. We can allow the expectation to be infinite, provided that we never get in a situation where we have '$\infty - \infty$'.

Expectation of $g(X)$

For a measurable function g that maps S_X to some other set S_Y, $Y = g(X)$ is a random variable taking values in S_Y and

$$E(Y) = E(g(X)) = \int_{S_X} g(x) dF_X(x).$$

We can also evaluate $E(Y)$ by calculating its distribution function $F_Y(y)$ and then using the expression

$$E(Y) = \int_{S_Y} y dF_Y(y).$$

Properties of Expectation

- $E(aX + bY) = aE(X) + bE(Y)$.
- If $X \leq Y$, then $E(X) \leq E(Y)$.
- If $X \equiv c$, then $E(X) = c$.

Moments

- The kth moment of X is $E(X^k)$.
- The kth central moment of X is $E[(X - E(X))^k]$.
- The variance $V(X)$ of X is the second central moment $E(X^2) - (E(X))^2$.
- $V(cX) = c^2 V(X)$.
- If X and Y have finite means and are independent, then $E(XY) = E(X)E(Y)$.
- If X and Y are independent (or uncorrelated), then $V(X \pm Y) = V(X) + V(Y)$.

Conditional Probability

The conditional probability of event A given X is a random variable (since it is a function of X). We write it as $P(A|X)$.

- for a real number x, if $P(X = x) > 0$, then $P(A|x) = P(A \cap \{X = x\})/P(\{X = x\})$.
- if $P(X = x) = 0$, then

$$P(A|x) = \lim_{\epsilon \to 0^+} P(A \cap \{X \in (x-\epsilon, x+\epsilon)\})/P(\{X \in (x-\epsilon, x+\epsilon)\}).$$

Conditional Distribution

- The conditional distribution function $F_{Y|X}(y|X)$ of Y evaluated at the real number y is given by $P(\{Y \leq y\}|X)$, where $P(\{Y \leq y\}|x)$ is defined on the previous slide.
- If (X, Y) is absolutely continuous, then the conditional density of Y given that $X = x$ is $f_{Y|X}(y|x) = f_{(X,Y)}(x, y)/f_X(x)$.

Conditional Expectation

The conditional expectation $E(Y|X) = \eta(X)$ where

$$\begin{aligned}
\eta(x) &= E(Y|X = x) \\
&= \begin{cases} \sum_j y_j P(Y = y_j|X = x) \text{ if } Y \text{ is discrete} \\ \int_{S_Y} y f_{Y|X}(y|x) dy \text{ if } Y \text{ is absolutely continuous.} \end{cases}
\end{aligned}$$

Properties of Conditional Expectation

- Linearity: $E(aY_1 + bY_2|X) = aE(Y_1|X) + bE(Y_2|X)$,
- Monotonicity: $Y_1 \leq Y_2$, then $E(Y_1|X) \leq E(Y_2|X)$,
- $E(c|X) = c$,
- $E(E(Y|X)) = E(Y)$,
- For any measurable function g, $E(g(X)Y|X) = g(X)E(Y|X)$
- $E(Y|X)$ is the best predictor of Y from X in the mean square sense. This means that, for all random variables $Z = g(X)$, the expected quadratic error $E((g(X) - Y)^2)$ is minimised when $g(X) = E(Y|X)$ (see Borovkov, page 57).

Exercise

Let $\Omega = \{a, b, c, d\}$, $P(\{a\}) = \frac{1}{2}$, $P(\{b\}) = P(\{c\}) = \frac{1}{8}$ and $P(\{d\}) = \frac{1}{4}$.
Define random variables,

$$
\begin{aligned}
Y(\omega) &= \begin{cases} 1, & \omega = a \text{ or } b, \\ 0, & \omega = c \text{ or } d, \end{cases} \\
X(\omega) &= \begin{cases} 2, & \omega = a \text{ or } c, \\ 5, & \omega = b \text{ or } d. \end{cases}
\end{aligned}
$$

Compute $E(X)$, $E(X|Y)$ and $E(E(X|Y))$.

Example

The number of storms, N, in the upcoming rainy season is distributed according to a Poisson distribution with a parameter value Λ that is itself random. Specifically, Λ is uniformly distributed over $(0, 5)$. The distribution of N is called a mixed Poisson distribution.

1. Find the probability there are at least two storms this season.
2. Calculate $E(N|\Lambda)$ and $E(N^2|\Lambda)$.
3. Derive the mean and variance of N.

Exercise

The joint density of X and Y is given by

$$f_{X,Y}(x,y) = \frac{e^{-x/y}e^{-y}}{y},\ x > 0,\ y > 0.$$

Calculate $E[X|Y]$ and then calculate $E[X]$.

Limit Theorems (Borovkov §2.9)

The Law of Large Numbers (LLN) states that if X_1, X_2, $\cdots$ are independent and identically-distributed with mean μ, then

$$\overline{X_n} \equiv \frac{1}{n}\sum_{j=1}^{n} X_j \to \mu$$

as $n \to \infty$.
In the strong form, this is true almost surely, which means that it is true on a set A of sequences $x_1, x_2, \ldots$ that has probability one.
In the weak form, this is true in probability which means that, for all $\epsilon > 0$,

$$P(|\overline{X_n} - \mu| > \epsilon) \to 0$$

as $n \to \infty$.

Limit Theorems (Borovkov §2.9)

The Central Limit Theorem (CLT) states that if X_1, X_2, $\cdots$ are independent and identically-distributed with mean μ and variance σ^2, then for any x,

$$P\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < x\right) \to \Phi(x) \equiv \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

as $n \to \infty$.
That is, a suitably-scaled variation from the mean approaches a standard normal distribution as $n \to \infty$.

Limit Theorems (Borovkov §2.9)

The Poisson Limit Theorem states that that if X_1, X_2, $\cdots$ are independent Bernoulli random variables with $P(X_i = 1) = 1 - P(X_i = 0) = p_i$, then $X_1 + X_2 + \cdots + X_n$ is well-approximated by a Poisson random variable with parameter $\lambda = p_1 + \cdots + p_n$.

Specifically, with $W = X_1 + X_2 + \cdots + X_n$, then, for any Borel set $B \subset R$,

$$P(W \in B) \approx P(Y \in B)$$

where $Y \sim \mathrm{Po}(\lambda)$.

There is, in fact, a bound on the accuracy of this approximation

$$|P(W \in B) - P(Y \in B)| \leq \frac{\sum_{i=1}^{n} p_i^2}{\max(1, \lambda)},$$

Example

Suppose there are three ethnic groups, A (20%), B (30%) and C (50%), living in a city with a large population. Suppose 0.5%, 1% and 1.5% of people in A, B and C respectively are over 200cm tall. If we know that of 300 selected, 50, 50 and 200 people are from A, B and C, what is the probability that at least four will be over 200 cm?

Stochastic Processes (Borovkov §2.10)

A collection of random variables $\{X_t,\ t \in T\}$ (or $\{X(t), t \in T\}$) on a common prob space $(\Omega, \mathcal{F}, P)$ is called a stochastic process. The index variable t is often called 'time'.

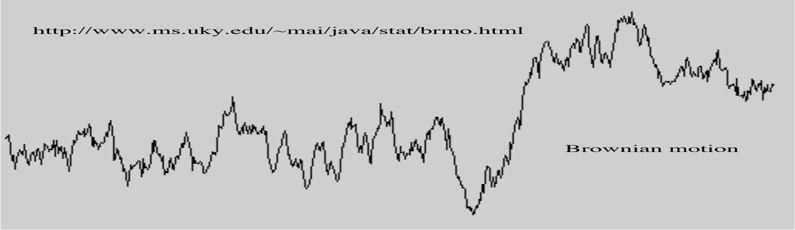
- If $T = \{1, 2, \cdots\}$ or $\{\cdots, -2, -1, 0, 1, 2, \cdots\}$, the process is a discrete time process.
- If $T = \mathbb{R}$ or $[0, \infty)$, the process is a continuous time process
- If $T = \mathbb{R}^d$, then the process is a spatial process, for example temperature at $t \in T \subset \mathbb{R}^2$, which could be, say, the University campus.

Examples of Stochastic Processes

If X_t has the normal distribution for all t, then X_t is called a *Gaussian* process. Different processes can be modelled by making different assumptions about the dependence between the X_t for different t.

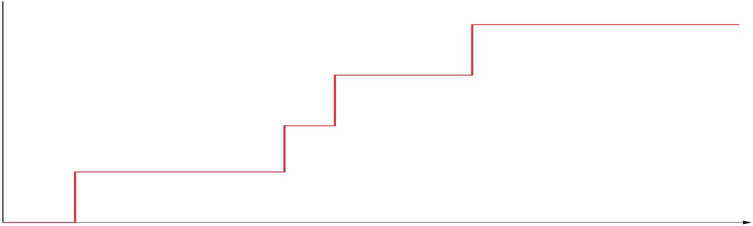
Standard Brownian Motion is a Gaussian process where

- For any $0 \leq s1 < t_1 \leq s_2 < \cdots \leq s_k < t_k$, $X(t_1) - X(s_1)$, $\ldots$, $X(t_k) - X(s_k)$ are independent.
- We also have $V(X(t_1) - X(s_1)) = t_1 - s_1$ for all $s_1 < t_1$.



Examples of Stochastic Processes

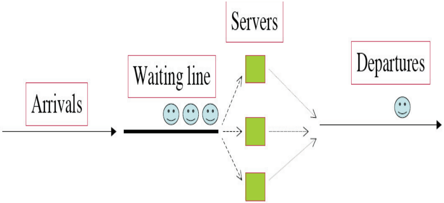
X_t is the number of sales of an item up to time t.

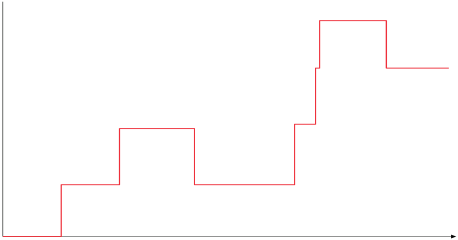


$\{X_t, t \geq 0\}$ is called a counting process.

Examples of Stochastic Processes

X_t is the number of people in a queue at time t.





Interpretations

We can think of Ω as consisting of the set of sample paths $\Omega = \{X_t : t \in T\}$, that is a set of sequences if T is discrete or a set of functions if T is continuous. Each $\omega \in \Omega$ has a value at each time point $t \in T$. With this interpretation,

- For a fixed ω, we can think of t as a variable, $X_t(\omega)$ as a deterministic function (realisation, trajectory, sample path) of the process.
- If we allow ω to vary, we get a collection of trajectories.
- For fixed t, with ω varying, we see that $X_t(\omega)$ is a random variable.
- If both ω and t are fixed, then $X_t(\omega)$ is a real number.

Examples of Stochastic Processes

If X_t is a counting process:

- For fixed ω, $X_t(\omega)$ is a non-decreasing step function of t.
- For fixed t, $X_t(\omega)$ is a non-negative integer-valued random variable.
- For $s < t$, $X_t - X_s$ is the number of events that have occurred in the interval $(s, t]$.

If X_t is the number of people in a queue at time t, then $\{X_t : t \geq 0\}$ is a stochastic process where, for each t, $X_t(\omega)$ is a non-negative integer-valued random variable but it is NOT a counting process because, for fixed ω, $X_t(\omega)$ can decrease.

Finite-Dimensional Distributions

Knowing just the one-dimensional (individual) distributions of X_t for all t is not enough to describe a stochastic process.

To specify the complete distribution of a stochastic process $\{X_t,\ t \in T\}$, we need to know the finite-dimensional distributions.

That is, the family of joint distribution functions

$$F_{t_1,t_2,\cdots,t_k}(x_1,\cdots,x_k)$$

of $X_{t_1},\cdots,X_{t_k}$ for all $k \geq 1$ and $t_1,\cdots,t_k \in T$.