Tutorial on Bayesian learning and related methods
A pre-seminar for Simon Godsill’s talk

Simon Wilson
Trinity College Dublin
Probability: the basics (1)

- Probability: the mathematics of describing uncertain quantities;
- \( P(A) \) is the probability that the event \( A \) occurs;
- Rules of probability are:
  1. \( 0 \leq P(A) \leq 1 \);
  2. If \( A \) and \( B \) are mutually exclusive then \( P(A \text{ or } B) = P(A) + P(B) \);
  3. \( P(A \text{ and } B) = P(A) \cdot P(B \mid A) \), where \( P(B \mid A) \) means the probability of \( B \) given that \( A \) has occurred.
- Two events \( A \) and \( B \) are called independent if \( P(B \mid A) = P(B) \), and so \( P(A \text{ and } B) = P(A) \cdot P(B) \).
A random variable is real-valued and its value is uncertain. Denoted by a capital letter e.g. $X$, its value by small letter $x$;

Random variables can be discrete or continuous;

If discrete, $X$ is described by its probability mass function, $p_X(x) = P(X = x)$;
- $p_X(x) \geq 0$, $\sum_{\forall x} p_X(x) = 1$.

If continuous, the probability density function $p_X(x)$ is used:
- $p_X(x) \geq 0$ and has the property that $P(a < X < b) = \int_a^b p_X(x) \, dx$;

The cumulative distribution function $F_X(x)$ is $P(X \leq x)$

$$F_X(x) = \begin{cases} \sum_{s \leq x} p_X(s), & \text{if } X \text{ discrete,} \\ \int_{s \leq x} p_X(s) \, ds, & \text{if } X \text{ continuous.} \end{cases}$$
The expected value or mean of a random variable is:

$$\mathbb{E}(X) = \begin{cases} \sum_{x} x p_X(x), & \text{if } X \text{ is discrete;} \\ \int_{\mathbb{X}} x p_X(x) \, dx, & \text{if } X \text{ is continuous.} \end{cases}$$

It’s the ’average’ value of $X$, the ’centre of gravity’ of the distribution;

The variance of $X$ is $\mathbb{E}((X - \mathbb{E}(X))^2)$:

$$\text{Var}(X) = \begin{cases} \sum_{x} (x - \mathbb{E}(X))^2 p_X(x), & \text{if } X \text{ is discrete;} \\ \int_{\mathbb{X}} (x - \mathbb{E}(X))^2 p_X(x) \, dx, & \text{if } X \text{ is continuous.} \end{cases}$$

It’s a measure of how variable the value of $X$ can be;

The standard deviation is $\sqrt{\text{Var}(X)}$;

- It has the same units of measurement as $X$ and $\mathbb{E}(X)$. 
Examples of discrete random variable distributions are the Bernoulli, binomial and Poisson:

\[
p_X(x \mid p) = p^x(1 - p)^{1-x}, \ x \in \{0, 1\};
\]

\[
p_X(x \mid n, p) = \binom{n}{x} p^x(1 - p)^{n-x}, \ x \in \{0, 1, \ldots, n\};
\]

\[
p_X(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \ x = 0, 1, 2, \ldots
\]

Are these familiar?

Good models for many physical phenomena;

Note that they are all defined in terms of parameters — \( p, n, \lambda \) — we think of these as conditional distributions of \( X \) given the parameter.
Examples of continuous random variable distributions are the exponential and normal (or Gaussian):

\[ p_X(x \mid \mu) = \frac{1}{\mu} e^{-x/\mu} \quad x \geq 0; \]

\[ p_X(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right), \quad x \in \mathbb{R}. \]

Are these familiar?

The normal distribution occurs in many places (and will in the seminar, repeatedly)

- \(\mu\) is its mean and \(\sigma^2\) is its variance;
Some normal pdf plots

- \( N(0,1) \)
- \( N(-1,1) \)
- \( N(0,4) \)
If we have two random variables $X$ and $Y$ then we can define the joint pmf/pdf $p(x, y)$:

- For discrete $X$, $Y$, $p(x, y) = P(X = x \text{ and } Y = y)$;
- For continuous $X$, $Y$,
  \[
  \int_a^b \int_c^d p(x, y) \, dx \, dy = P(c < X < d, a < Y < b);
  \]
- The laws of probability show that $p_X(x) = \int_{\forall y} p(x, y) \, dy$;
- For discrete $X$ and $Y$, the \textit{conditional distribution} of $X$ given $Y = y$ is
  \[
  P(X = x \mid Y = y) = \frac{p(x, y)}{p_Y(y)}.
  \]

$X$ and $Y$ are called \textit{independent} if

- $P(X = x \mid Y = y) = P(X = x)$ and
- $P(Y = y \mid X = x) = P(Y = y)$;

In this case, $p(x, y) = p_X(x) \, p_Y(y)$.
Two urns: urn I has 3 red and 3 blue balls, urn II has 2 red and 4 green balls;
I flip a fair coin (so $P(H) = P(T) = 1/2$). If $H$ then pick a ball from urn I else pick one from urn II.
What is $P(R) = P(\text{red ball picked})$? By laws of probability:

$$P(R) = P((H \text{ and } R) \text{ or } (T \text{ and } R))$$
$$= P(H \text{ and } R) + P(T \text{ and } R)$$
$$= P(H) P(R | H) + P(T) P(R | T)$$
$$= \sum_{y=H,T} P(y) P(R | y)$$
$$= 1/2 \times 3/6 + 1/2 \times 2/6 = 5/12.$$
Two urns: urn I has 3 red and 3 blue balls, urn II has 2 red and 4 green balls;

I flip a fair coin. If $H$ then pick a ball from urn I else pick one from urn II.

Now I tell you that I picked a red ball. What is the chance that I flipped a $H$? This is $P(H \mid R)$:

$$P(H \mid R) = \frac{P(H \text{ and } R)}{P(R)} = \frac{P(H) P(R \mid H)}{P(R)} = \frac{(1/2 \times 3/6)}{(5/12)} = \frac{3}{5}.$$

Note that if I do not tell you the ball colour, the chance of a $H$ is $P(H) = 1/2$.

Observing $R$ has allowed you to learn about how likely is $H$. 
The first equation is an example of the *partition law*.

In terms of random variables, we write that for any two random variables $X$ and $Y$:

$$p_X(x) = \sum_{y} p_{X|Y}(x \mid Y = y) \ p_Y(y),$$

where $p_{X|Y}(x \mid Y = y)$ is the *conditional pmf* of $X$ given $Y$. 
The second equation is an example of Bayes’ law.

It can be written as:

\[ p_{Y|X}(y|X = x) = \frac{p_Y(y) p_{X|Y}(x|Y = y)}{p_X(x)}, \]

which is often written (by Partition law):

\[ p_{Y|X}(y|X = x) = \frac{p_Y(y) p_{X|Y}(x|Y = y)}{\sum_{\forall y} p_Y(y) p_{X|Y}(x|Y = y)}. \]

We also see Bayes law written as:

\[ p_{Y|X}(y|X = x) \propto p_Y(y) p_{X|Y}(x|Y = y). \]

\( \sum \) replaced by \( \int \) if \( Y \) is continuous.
What is Monte Carlo simulation?

- This is to generate a sequence of values from a probability distribution;
- Usually done by computer;
- To Monte Carlo simulate from a probability distribution $p_X(x)$ means to generate a sequence of values $x_1, x_2, \ldots, x_N$ such that:
  - The values are independent
  - If $X$ discrete, the proportion of values equal to $x$ converges to $p_X(x)$, $\forall x$ as $N \to \infty$;
  - If $X$ is continuous, the proportion of values in the interval $(a, b)$ converges to \( \int_a^b p_X(x) \, dx \), $\forall a, b$ as $N \to \infty$;
- E.g. simulation of a die: 4, 2, 5, 5, 1, 6, 3, 4, 2, 1, \ldots
- What about 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, \ldots?
Pseudo-random numbers

- Computers are logic machines ⇒ should be no good at Monte Carlo simulation!
- This is true — the best we can do is generate deterministic sequences of numbers;
  - These sequences have many of the properties of 'really' random sequences;
  - For most purposes they are indistinguishable from using 'the real thing';
  - They can also be generated very quickly ($\approx 10^7$ / second);
- The basis of Monte Carlo methods are random numbers — these are uniformly distributed between 0 and 1 ($p_X(x) = 1, 0 \leq x \leq 1$);
- There are many algorithms for generating deterministic sequences that look like random numbers:
  - These are called pseudo-random numbers;
There are many methods of generating values from other probability distributions: discrete, normal, exponential, etc;

All of these rely on a supply of pseudo-random numbers;

Many computer packages are able to Monte Carlo simulate from many distributions: MATLAB, R, even Excel!
Example: discrete probability distributions

- Let \( X \) be tomorrow’s weather, \( X \in \{ \text{sunny, cloudy, rainy} \} \);
- Suppose \( P(X = S) = 0.2, P(X = C) = 0.3, P(X = R) = 0.5; \)
- We can Monte Carlo simulate this distribution as follows:
  - Generate a (pseudo-) random number \( u \);
  - If \( u < 0.2 \) then \( X = S \);
  - if \( 0.2 \leq u < 0.5 \) then \( X = C \);
  - if \( u \geq 0.5 \) then \( X = R \).
- This idea is called the inverse distribution method and works for all discrete distributions.
Bayes law is the basis for learning

- In the urn problem, observing $R$ tells you something about the coin flip but does not tell you if it’s $H$ or $T$ with certainty;
- The question is then: how “certain” can I be that the flip is a $H$? Or $T$?
- Bayes’ law allowed us to compute how certain, as a probability, in terms of probabilities that we know.
- This situation occurs everywhere in data analysis and is the basis of statistical inference (or statistical learning);
- Bayesian statistical inference defines what we learn through a probability distribution on the quantity of interest;
- Often this is defined through Bayes’ law
What is the temperature in this room?

For simplicity, let’s assume that it’s constant all over the room.

I have a thermometer and it measures $18.1^\circ C$;

- Is that the “real” temperature in the room?
- Why not?

I have another identical make of thermometer. It measures $18.1^\circ C$ as well.

- Should I be more certain about the value of the real temperature now?
- If yes then by how much?
- What if the second thermometer had read $18.4^\circ C$?
What does Bayesian inference say about how to answer this question?

Let $T$ be the true (and unknown) temperature in the room;

Let $x_1$ and $x_2$ be the temperature measurements;

Our state of knowledge about $T$ is defined by $p_T(t \mid x_1, x_2)$;

By Bayes’ law:

$$p_T(t \mid x_1) = \frac{p_T(t) p(x_1 \mid T)}{p(x_1)} \propto p_T(t) p(x_1 \mid T);$$

$$p_T(t \mid x_1, x_2) \propto p_T(t) p(x_1, x_2 \mid T).$$
Slightly more complicated example (3)

- $p_T(t)$ represents what we think $T$ is before we measure it;
- This is known as the prior distribution;
- For example, we are pretty sure that $0 \leq T \leq 40$; one possibility is a uniform distribution on this range:

\[
p_T(t) = \frac{1}{40}, \quad 0 \leq t \leq 40.
\]

- Another is a normal distribution with mean as our best guess (say $20^\circ C$) and a standard deviation of 10 (so that $0 \leq T \leq 40$ with high probability)
Two possible priors for $T$

![Possible Prior Distributions](image)

- **Uniform**
- **Normal**
Slightly more complicated example (4)

- $p(x_1, x_2 \mid T = t)$ describes what we measure given the true temperature is $t$;
- One reasonable model is that what we measure is normally distributed with mean $T$ and a variance $\sigma^2$;
- Here we assume that we know $\sigma^2$ (it says how accurate our thermometer is — let’s say $\sigma^2 = 0.3^2$);
- 
  $$p(x_1 \mid T = t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-0.5(x_1-t)^2/\sigma^2}.$$  
- Also assume that the two measurements are *independent* given $T$, so that:

  $$p(x_1, x_2 \mid T = t) = p(x_1 \mid T = t) p(x_2 \mid t = t)$$
  $$= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2} [(x_1-t)^2+(x_2-t)^2]}.$$
Distribution of $x_1$ when $T = 18$
In Bayes' law, the variable is $t$, so we should actually think of $p(x_1 \mid T = t)$ and $p(x_1, x_2 \mid T = t)$ as a function of $t$;

This is called the *likelihood*; with $\sigma^2 = 0.3^2$ we have:

$$p(x_1 \mid T = t) = \frac{1}{\sqrt{0.18\pi}} e^{-(x_1-t)^2/0.18}.$$ 

Also assume that the two measurements are *independent* given $T$, so that:

$$p(x_1, x_2 \mid T = t) = p(x_1 \mid T = t) p(x_2 \mid t = t)$$

$$= \frac{1}{0.18\pi} e^{-\frac{1}{0.18}[(x_1-t)^2+(x_2-t)^2]}.$$
Likelihood for $x_1 = 18.1$
Likelihood for $x_1 = 18.1$, $x_2 = 18.1$
Likelihood for $x_1 = 18.1$, $x_2 = 18.4$
Bayes’ law gave us:

\[ p_T(t \mid x_1, x_2) \propto p_T(t) \times p(x_1, x_2 \mid T = t); \]

- \( p_T(t \mid x_1, x_2) \) is called the *posterior distribution*
- Plots of \( p_T(t) \times p(x_1, x_2 \mid T = t) \) next.
\[ p_T(t) \times p(x_1 = 18.1 \mid T = t) \quad \text{and} \quad p_T(t) \times p(x_1 = x_2 = 18.1 \mid T = t) \]
\( p_T(t) \times p(x_1 = x_2 = 18.1 \mid T = t) \) and 
\( p_T(t) \times p(x_1 = 18.1, x_2 = 18.4 \mid T = t) \)
Now all that is missing is the constant that relates $p_T(t \mid x_1, x_2)$ to $p_T(t) \times p(x_1, x_2 \mid T = t)$;

Bayes law:

$$p_T(t \mid x_1, x_2) = \frac{p_T(t) \times p(x_1, x_2 \mid T = t)}{p(x_1, x_2)}$$

tells you that this is

$$1/p(x_1, x_2) = 1/\int_0^\infty p_T(t) \times p(x_1, x_2 \mid T = t) \, dt$$

It’s just the integral of the function plotted on the last slides;

It’s there to ensure that $\int_0^\infty p_T(t \mid x_1, x_2) \, dt = 1$;

$1/p(18.1, 18.1) = 55.6$, $1/p(18.1, 18.4) = 68.7$
\( p(t \mid x_1 = x_2 = 18.1) \) and \( p(t \mid x_1 = 18.1, x_2 = 18.4) \)
Stochastic processes

Many of the processes that we want to model occur over space and time:

- Audio signals;
- Rainfall;
- Financial time series
- Image and video data...

These are modelled probabilistically by *stochastic processes*
A very common subset is processes that evolve at discrete points in time \( t = 1, 2, 3, \ldots, T \);

Then \( X_1, X_2, \ldots, X_T \) are the values of the process at these times;

In general, we then have to define a probability distribution on \((X_1, \ldots, X_T)\);

This is in general difficult because we have to define a \( T \)-dimensional function \( p(x_1, \ldots, x_T) \).

We can exploit properties of the process to make the model simpler to define:

Many processes obey what is called the *Markov property*;

This means that the distribution of \( X_t \) only depends on the value of \( X_{t-1} \);

If \( X_1, X_2, \ldots \) obeys this property then it’s called a *Markov chain*. 
Markov chains

- For a Markov chain:

\[ p(x_1, \ldots, x_T) = p(x_1) \cdot p(x_2 | x_1) \cdot p(x_3 | x_2) \cdots p(x_T | x_{T-1}). \]

- So \( p(x_1, \ldots, x_T) \) is defined in terms of simple one-dimensional distributions;

- If \( p(x_t | x_{t-1}) \) independent of \( t \) then we just need to define \( p(x_1) \) and \( p(x_t | x_{t-1}) \);

- If \( x_t \) is discrete-valued (say \( x_t \in \{1, 2, \ldots, S\} \)) then \( p(x_t | x_{t-1}) \) defined in a matrix:

\[
P = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1S} \\
p_{21} & p_{22} & \cdots & p_{2S} \\
\vdots & \vdots & \ddots & \vdots \\
p_{S1} & p_{S2} & \cdots & p_{SS}
\end{pmatrix},
\]

where \( p_{ij} = P(X_t = j | X_{t-1} = i) \). Hence each row in \( P \) sums to 1.
The weather each day is: sunny (S), cloudy (C) or rainy (R);

$X_t$ is the weather on day $t$;

Suppose the weather on day $t$ depends on the weather on day $t - 1$, but is independent of earlier days;

It is then a Markov chain and suppose

\[
P = \begin{pmatrix}
0.3 & 0.5 & 0.2 \\
0.25 & 0.5 & 0.25 \\
0.4 & 0.3 & 0.3
\end{pmatrix}
\]

\[
\begin{array}{c}
S \\
C \\
R
\end{array}
\]

\[
x_{t-1}
\]

e.g. $P(x_t = S \mid x_{t-1} = R) = p_{31} = 0.4;
Monte Carlo simulation of our Markov chain

- This is easy to do;
- Define $x_1$ (let’s make it $x_1 = R$);
- Simulate $x_2$ given $x_1$ (using the $R$ row of $P$) — suppose we generate $x_2 = C$;
- Simulate $x_3$ given $x_2$ (using the $C$ row of $P$);
- and so on.
40 days of weather

Sunny = 1, Cloudy = 2, Rainy = 3

\[ x_t \]

\[ t \]

5 10 15 20 25 30 35 40
A random walk

- Markov chains can have $X_t$ continuous;
- Example: $X_t$ is normally distributed with mean $X_{t-1}$ and a variance $\sigma^2$;
- This is known as a random walk;
- On next page is a simulation with $X_1 = 0$ and two values of $\sigma^2$
A normal random walk
Autoregressive processes

- Random walks have the property that they 'wander away' from 0 (they are *non-stationary*);
- Many physical processes tend to stay around a mean value (they are *stationary*);
- An *autoregressive process* is a simple case: $X_t$ is normally distributed with mean $\theta X_{t-1}$ and a variance $\sigma^2$, where $-1 < \theta < 1$;
- Higher order autoregressive processes are also very common e.g. $X_t$ has mean $\theta_1 X_{t-1} + \theta_2 X_{t-2}$, etc.;
- These are a simple model for an audio signal.
First order autoregressive processes (with $\sigma = 1$)

$$X_t$$

$\theta = 0.7$

$\theta = 0.1$
Third order autoregressive process

\[ E(X_t) = 0.7X_{t-1} + 0.1X_{t-2} + 0.15X_{t-2} \]
Properties of Markov chains

- Vast literature in probability theory on properties of Markov chains;
- Here, we concentrate on one property;
- Suppose I start our weather chain in day 1;
- What is the weather on day $t + 1$? This is $P(x_{t+1} \mid x_1)$;
- It turns out that the matrix of these probabilities is

$$P^t = P \times P \times \cdots \times P \text{ (matrix multiplication)}.$$
Properties of Markov chains

Here is $P^t$ for $t = 2, 20, 200$:

$$
\begin{pmatrix}
0.30 & 0.46 & 0.24 \\
0.30 & 0.45 & 0.25 \\
0.32 & 0.44 & 0.24
\end{pmatrix},
\begin{pmatrix}
0.30 & 0.45 & 0.25 \\
0.30 & 0.45 & 0.25 \\
0.30 & 0.45 & 0.25
\end{pmatrix},
\begin{pmatrix}
0.30 & 0.45 & 0.25 \\
0.30 & 0.45 & 0.25 \\
0.30 & 0.45 & 0.25
\end{pmatrix}.
$$
Stationary distributions

- So as you look further into the future:
  - The probability that you are in each state converges to a value;
  - This probability is the \textit{same} regardless of your current state (so the chain ‘forgets’ where it was at the start);
  - This is called the \textit{stationary distribution} of the chain (in this case \((0.30, 0.45, 0.25)\));
  - Not all Markov chains have such a distribution;
  - Lots of theory on conditions for which it does happen — includes most chains that one uses in practice;
Simulating the weather Markov chain — proportion of sunny days
Markov chains for Monte Carlo simulation

- Note that the proportion of sunny days in the simulation converges to the stationary probability of 0.302;
- So there is (very long winded!!) way to simulate from the distribution \((0.30, 0.45, 0.25)\):
  - Start this Markov chain in any of the 3 states;
  - Monte Carlo simulate the chain for a 'long' time;
  - The state of the chain at the end of the long simulation will have distribution \((0.30, 0.45, 0.25)\);
- Hold this thought for later!
Audio reconstruction

- Think of an audio signal as a discrete process in time $X_1, X_2, \ldots$;
- We have some audio data that has been corrupted:
  - CD or vinyl record scratches, tape degradation;
  - Telephone call over a noisy line;
- So what we observe are not the $X_t$ but a corrupted version $Y_1, \ldots, Y_T$;
- We want to recover the original 'true' audio signal $X_1, \ldots, X_T$ from $Y_1, \ldots, Y_T$;
Hidden Markov models

- A simple model for this process is as follows:
  - Our real audio signal is a stochastic process like the AR model;
  - What we actually observe $Y_t$ is normally distributed with mean $X_t$;
  - The normal distribution models the noise or degradation of the signal’
  - The $Y_t$’s are independent of each other given the $X_t$’s;

- Mathematically:

  \[
  Y_t \mid X_t, \sigma_Y^2 \sim N(X_t, \sigma_Y^2) \\
  X_t \mid X_{t-1}, \sigma_X^2 \sim N(\theta X_{t-1}, \sigma_X^2).
  \]

- This is an example of a Hidden Markov Model (HMM):
  - There is hidden Markov chain $X_1, X_2, \ldots$;
  - We observe $Y_t$ that are $X_t$ plus some noise, and are independent;
$X_t$ – an AR process
$X_t$ and $Y_t$
$Y_t$ alone
Bayesian inference (1)

- Our data are the $Y_t$'s;
- The likelihood is:

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_t, \sigma_Y^2) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left( - \frac{1}{2\sigma_Y^2} (y_t - x_t)^2 \right).$$

- We want to learn about the $X_t$ and the 3 parameters ($\sigma_Y^2, \theta, \sigma_X^2$);
Bayesian inference (2)

- The prior distribution for the $X_t$ is (say) an AR process:

$$p(x_1, \ldots, x_t | \theta, \sigma^2_X) = \frac{1}{\sqrt{2\pi \sigma^2_X}} e^{-x_1^2/2\sigma^2_X} \prod_{i=2}^{T} \frac{1}{\sqrt{2\pi \sigma^2_X}} e^{-(x_t - \theta x_{t-1})^2/2\sigma^2_X}.$$ 

- We have a prior for the 3 parameters $p(\sigma_Y^2, \theta, \sigma^2_X)$;

- Then Bayes’ law gives us:

$$p(x_1, \ldots, x_T, \sigma^2_X, \theta, \sigma_Y^2 | y_1, \ldots, y_T) = \frac{p(y_1, \ldots, y_T | x_1, \ldots, x_t, \sigma_Y^2) p(x_1, \ldots, x_t | \theta, \sigma^2_X) p(\sigma_Y^2, \theta, \sigma^2_X)}{p(y_1, \ldots, y_T)}$$
What is the denominator?

\[
p(y_1, \ldots, y_T) = \int p(y_1, \ldots, y_T \mid x_1, \ldots, x_t, \sigma_Y^2) p(x_1, \ldots, x_t \mid \theta, \sigma_X^2) \times p(\sigma_Y^2, \theta, \sigma_X^2),
\]

a \(T + 3\) dimensional integral;

This is in general impossible to compute numerically — too big a problem!

What can we do?

- For this problem, there are some algorithms that can compute the means of the \(X_i\), and approximate the posterior;
- These break down when we consider more realistic models for \(X_t\) and \(Y_t \mid X_t\);
- Can we Monte Carlo simulate from \(p(x_1, \ldots, x_T, \sigma_X^2, \theta, \sigma_Y^2 \mid y_1, \ldots, y_T)\)?
MC simulation from high-dimensional distributions is also very difficult;

However, it is possible to simulate from a Markov chain with stationary distribution that is the posterior;

So we simulate values of \((x_1, \ldots, x_T, \sigma_X^2, \theta, \sigma_Y^2)\) according to a certain Markov chain;

After we simulate values for a 'long time', we are sampling from the posterior.

For many Bayesian learning problems, this is the only way that we know to approximate the solution;

This is known as MCMC (Markov chain Monte Carlo);

MCMC methods are a module in themselves!
I hope that this has given you an idea of some of the methods used for Bayesian inference;

Remember: seminar is on Friday at 12pm in this room.