Housekeeping

• **Website:** www.stats.ox.ac.uk/~deligian/teaching/

• **Email:** deligian@stats.ox.ac.uk

• **Lectures:** Mondays 16-17:00 & Tuesdays 11-12:00, weeks 1-8, LG01.

• **UG Classes:**
  • Wednesday 9:30-11:00 @ LG04, weeks 2,5,7 and TT1; Tutor: Alex Shestopaloff;
  • Thursdays 09:00-10:30 @ LG04, weeks 2,5,7 and TT1; Tutor: James Thornton;
  • Thursdays 10:30-12:00 @ LG04, weeks 2,5,7 and TT1; Tutor: Soufiane Hayou;

• **MSc Classes:** Wednesdays 10:00-11:00 weeks 3, 5, 7, 8, @ LG01.

• Hand in solutions by Monday, 9am at the Adv. Simulation tray.
Motivation

Solutions of many scientific problems involve intractable high-dimensional integrals.

Standard deterministic numerical integration deteriorates rapidly with dimension.

Monte Carlo methods are stochastic numerical methods to approximate high-dimensional integrals.

Main application in this course: Bayesian statistics.

Other applications: statistical/quantum physics, econometrics, ecology, epidemiology, finance, signal processing, weather forecasting. . . , and machine learning.

As of today 3,700,000 results for “Monte Carlo” in Google Scholar.
Computing Integrals

For $f: \mathbb{X} \to \mathbb{R}$, let

$$I = \int_{\mathbb{X}} f(x) \, dx.$$

When $\mathbb{X} = [0, 1]$, then we can simply approximate $I$ through

$$\hat{I}_n = \frac{1}{n} \sum_{i=0}^{n-1} f \left( \frac{i + 1/2}{n} \right).$$
Figure: Riemann sum approximation (black rectangles) of the integral of $f$ (red curve).
Error of naive numerical integration in 1D

Naively, for a small interval \([a, a + \varepsilon]\) approximate

\[
\int_{a}^{a+\varepsilon} f(x) \, dx \approx \varepsilon \times f(a).
\]

Error bounded above by

\[
\left| \int_{a}^{a+\varepsilon} f(x) \, dx - \varepsilon \times f(a) \right| = \left| \int_{a}^{a+\varepsilon} [f(x) - f(a)] \, dx \right|
\]

\[
\leq \int_{a}^{a+\varepsilon} \int_{y=a}^{x} |f'(y)| \, dy \, dx \leq \sup_{x \in [0,1]} |f'(x)| \frac{\varepsilon^2}{2}.
\]

If \(\sup_{x \in [0,1]} |f'(x)| < M\), the uniform grid with \(n\) points gives approximation error at most

\[
Mn \times \frac{1}{n^2} = O\left(\frac{1}{n}\right).
\]
Computing High-Dimensional Integrals

For $\Xi = [0, 1] \times [0, 1]$ using $n = m^2$ evaluations

$$\hat{I}_n = \frac{1}{m^2} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} f \left( \frac{i + 1/2}{m}, \frac{j + 1/2}{m} \right)$$

the same calculation shows that the approximation error is

$$Mm^2 \times \frac{1}{m^3} = \mathcal{O}(1/m) = \mathcal{O} \left( n^{-1/2} \right).$$

Generally for $\Xi = [0, 1]^d$ we have an approximation error in

$$\mathcal{O} \left( n^{-1/d} \right).$$

So-called “curse of dimensionality”.

Other integration rules (e.g. Simpson’s) also degrade as $d$ increases.
Monte Carlo Integration

For \( f : \mathbb{X} \rightarrow \mathbb{R} \), write

\[
I = \int_{\mathbb{X}} f(x) \, dx = \int_{\mathbb{X}} \varphi(x) \pi(x) \, dx.
\]

where \( \pi \) is a probability density function on \( \mathbb{X} \) and

\[
\varphi : x \mapsto f(x) / \pi(x).
\]

Monte Carlo method:

sample \( X_1, \ldots, X_n \) i.i.d. \( \sim \pi \),
compute

\[
\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i).
\]

Strong law of large numbers: \( \hat{I}_n \rightarrow I \) almost surely;
Central limit theorem: the random approximation error is

\[
O(n^{-1/2})
\]

whatever the dimension of the state space \( \mathbb{X} \).
Computing High-Dimensional Integrals

Non-asymptotically, we can prove this result using the mean-square error. We have:

\[
(I - \hat{I}_n)^2 = I^2 - 2I \times \hat{I}_n + \hat{I}_n^2
\]

\[
= I^2 - \frac{2I}{n} \sum_{i=1}^{n} \varphi(X_i) + \frac{1}{n^2} \sum_{i=1}^{n} \varphi(X_i)^2 + \frac{1}{n^2} \sum_{i \neq j} \varphi(X_i) \varphi(X_j).
\]

As the samples are i.i.d. and \( I = \mathbb{E}_\pi[\varphi(X)] \), we have

\[
\mathbb{E}_\pi[(I - \hat{I}_n)^2] = I^2 - 2I^2 + \frac{1}{n} \mathbb{E}_\pi[\varphi(X_1)^2] + \frac{1}{n^2} n(n - 1)I^2
\]

\[
= \frac{\mathbb{E}_\pi[\varphi(X_1)^2] - I^2}{n} = \frac{\mathbb{V}_\pi(\varphi(X_1))}{n}
\]

and \( \sqrt{\mathbb{E}_\pi[(I - \hat{I}_n)^2]} = \frac{\sqrt{\mathbb{V}_\pi(\varphi(X_1))}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}} \) if \( |\varphi(x)| \leq 1 \ \forall x \).

The constant on the r.h.s. of the bound is 1, hence independent of the dimension of the state space \( \mathcal{X} \).
Monte Carlo Integration

In many cases the integral of interest is in the form

\[ I = \int_{\mathcal{X}} \varphi(x)\pi(x)dx = \mathbb{E}_\pi[\varphi(X)], \]

for a specific function \( \varphi \) and distribution \( \pi \).

The distribution \( \pi \) is often called the “target distribution”.

Monte Carlo approach relies on independent copies of

\[ X \sim \pi. \]

Hence the following relationship between integrals and sampling:

Monte Carlo method to approximate \( \mathbb{E}_\pi[\varphi(X)] \)
\[ \Leftrightarrow \] simulation method to sample \( \pi \)

Thus Monte Carlo sometimes refers to simulation methods.
Ising Model

Consider a simple 2D-Ising model

♣ a finite lattice \( \mathcal{G} = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m\} \)
♣ each site \( \sigma = (i,j) \) hosts a particle with a +1 or -1 spin modeled as a r.v. \( X_\sigma \).

The distribution of \( X = \{X_\sigma\}_{\sigma \in \mathcal{G}} \) on \( \{-1, 1\}^{m^2} \) is given by

\[
\pi_\beta(x) = \frac{\exp(-\beta \times U(x))}{Z_\beta}
\]

where \( \beta > 0 \) is called the inverse temperature and the potential energy is

\[
U(x) = J \sum_{\sigma \sim \sigma'} x_\sigma x_{\sigma'}.
\]

Physicists are interested in computing \( \mathbb{E}_{\pi_\beta}[U(X)] \) and \( Z_\beta \).

The dimension is \( m^2 \), where \( m \) can easily be \( 10^3 \).
Ising Model

Figure: One draw from the Ising model on a $500 \times 500$ lattice.
Option Pricing

Let $S(t)$ denote the price of a stock at time $t$.

**European option:** grants the holder the right to buy the stock at a fixed price $K$ at a fixed time $T$ in the future; the current time being $t = 0$.

At time $T$ the holder achieves a payoff of

$$\max\{S_T - K, 0\}.$$ 

With interest rate $r$, the expected discounted value at $t = 0$ is

$$\exp(-rT) \mathbb{E}[\max(0, S(T) - K)].$$
Option Pricing

If we knew explicitly the distribution of $S(T)$ then $\mathbb{E} \left[ \max(0, S(T) - K) \right]$ is a low-dimensional integral.

**Problem:** We only have access to a complex stochastic model for
\{$S(t)\}_{t \in \mathbb{N}}$

\[
S(t + 1) = g(S(t), W(t + 1)) \\
= g(g(S(t - 1), W(t)), W(t + 1)) \\
=: g^{t+1}(S(0), W(1), ..., W(t + 1))
\]

where \{$W(t)\}_{t \in \mathbb{N}}$ is a sequence of random variables and $g$ is a known function.
Option Pricing

The price of the option involves an integral over the $T$ latent variables

$$\{W(t)\}_{t=1}^T.$$ 

Assume these are independent with probability density function $p_W$.

We can write

$$\mathbb{E} \left[ \max \left( 0, S(T) - K \right) \right]$$

$$= \int \max \left[ 0, g^T(s(0), w(1), \ldots, w(T)) - K \right]$$

$$\times \left\{ \prod_{t=1}^T p_W(w(t)) \right\} dw(1) \cdots dw(T),$$

a high-dimensional integral.
Bayesian Inference

Given $\theta \in \Theta$, we assume that $Y$ follows a probability density function $p_Y(y; \theta)$.

Having observed $Y = y$, we want to perform inference about $\theta$.

In the frequentist approach $\theta$ is unknown but fixed; inference in this context can be performed based on

$$\ell(\theta) = \log p_Y(y; \theta).$$

In the Bayesian approach, the unknown parameter is regarded as a random variable $\vartheta$ and assigned a prior $p_\vartheta(\theta)$. 

Frequentist vs Bayesian

Probabilities refer to limiting relative frequencies. They are (supposed to be) objective properties of the real world.

Parameters are fixed unknown constants. Because they are not random, we cannot make any probability statements about parameters.

Statistical procedures should have well-defined long-run properties. For example, a 95% confidence interval should include the true value of the parameter with limiting frequency at least 95%.
Frequentist vs Bayesian

Probability describes degrees of subjective belief, not limiting frequency.

We can make probability statements about parameters, e.g.

\[ \mathbb{P}(\theta \in [-1, 1] \mid Y = y) \]

Observations produce a new probability distribution for the parameter, the posterior.

Point estimates and interval estimates may then be extracted from this distribution.
Bayesian Inference

Bayesian inference relies on the posterior

$$p_{\theta|Y}(\theta|y) = \frac{p_Y(y; \theta) p_{\theta}(\theta)}{p_Y(y)}$$

where

$$p_Y(y) = \int_{\Theta} p_Y(y; \theta) p_{\theta}(\theta) \, d\theta$$

is the so-called marginal likelihood or evidence.

Point estimates, e.g. posterior mean of $\theta$

$$\mathbb{E}(\theta|y) = \int_{\Theta} \theta \, p_{\theta|Y}(\theta|y) \, d\theta$$

can be computed.
Bayesian Inference

Credible intervals: an interval $C$ such that

$$P(\vartheta \in C|y) = 1 - \alpha.$$  

Assume the observations are independent given $\vartheta = \theta$ then the predictive density of a new observation $Y_{\text{new}}$ having observed $Y = y$ is

$$p_{Y_{\text{new}}|Y}(y_{\text{new}}|y) = \int_{\Theta} p_{Y}(y_{\text{new}}; \theta) p_{\vartheta|Y}(\theta|y) d\theta$$

Above predictive density takes into account the uncertainty about the parameter $\theta$.

Compare to simple plug-in rule $p_Y(y_{\text{new}}; \hat{\theta})$ where $\hat{\theta}$ is a point estimate of $\theta$ (e.g. the MLE).
Bayesian Inference: Gaussian Data

Let \( \mathbf{Y} = (Y_1, \ldots, Y_n) \) be i.i.d. random variables with \( Y_i \sim \mathcal{N}(\theta, \sigma^2) \) with \( \sigma^2 \) known and \( \theta \) unknown.

Assign a prior distribution on the parameter: \( \vartheta \sim \mathcal{N}(\mu, \kappa^2) \), then one can check that

\[
p(\theta|\mathbf{y}) = \mathcal{N}(\theta; \nu, \omega^2)
\]

where

\[
\omega^2 = \frac{\kappa^2 \sigma^2}{n\kappa^2 + \sigma^2}, \quad \nu = \frac{\sigma^2}{n\kappa^2 + \sigma^2} \mu + \frac{n\kappa^2}{n\kappa^2 + \sigma^2} \bar{y}.
\]

Thus \( \mathbb{E}(\vartheta|\mathbf{y}) = \nu \) and \( \mathbb{V}(\vartheta|\mathbf{y}) = \omega^2 \).
Bayesian Inference: Gaussian Data

If \( C := (\nu - \Phi^{-1} (1 - \alpha/2) \omega, \nu + \Phi^{-1} (1 - \alpha/2) \omega) \), then

\[
P(\theta \in C | y) = 1 - \alpha.
\]

If \( Y_{n+1} \sim \mathcal{N}(\theta, \sigma^2) \) then

\[
p(y_{n+1} | y) = \int_{\Theta} p(y_{n+1} | \theta) p(\theta | y) \, d\theta = \mathcal{N}(y_{n+1}; \nu, \omega^2 + \sigma^2).
\]

No need to do Monte Carlo approximations: the prior is conjugate for the model.
Bayesian Inference: Logistic Regression

Let \((x_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}\) where \(x_i \in \mathbb{R}^d\) is a covariate and

\[
P(Y_i = 1 | \theta) = \frac{1}{1 + e^{-\theta^T x_i}}
\]

Assign a prior \(p(\theta)\) on \(\vartheta\). Then Bayesian inference relies on

\[
p(\theta | y_1, \ldots, y_n) = \frac{p(\theta) \prod_{i=1}^{n} P(Y_i = y_i | \theta)}{P(y_1, \ldots, y_n)}
\]

If the prior is Gaussian, the posterior is not a standard distribution: \(P(y_1, \ldots, y_n)\) cannot be computed.
S&P 500 index

Figure: S&P 500 daily price index ($p_t$) between 1984 and 1991.
S&P 500 index

Figure: Daily returns $y_t = \log(p_t/p_{t-1})$ between 1984 and 1991.
Bayesian Inference: Stochastic Volatility Model

Latent stochastic volatility \((X_t)_{t \geq 1}\) of an asset is modeled through

\[
X_t = \varphi X_{t-1} + \sigma V_t, \quad Y_t = \beta \exp(X_t) W_t
\]

where \(V_t, W_t \sim \mathcal{N}(0, 1)\).

Intuitively, log-returns are modeled as centered Gaussians with dependent variances.

Popular alternative to ARCH and GARCH models (Engle, 2003 Nobel Prize).

Estimate the parameters \((\varphi, \sigma, \beta)\) given the observations.

Estimate \(X_t\) given \(Y_1, ..., Y_t\) on-line based on \(p(x_t | y_1, ..., y_t)\).

No analytical solution available!