3.36pt

Advanced Simulation - Lecture 1

George Deligiannidis

January 18th, 2016

Housekeeping

- First half of course: GD, second half: Lawrence Murray
- Website: `www.stats.ox.ac.uk/~deligian/sc5.html`
- Email: `deligian@stats.ox.ac.uk`
- **Lectures**: Mondays 10-11 & Wednesdays 14-15, weeks 1-8, LG01.
- **Classes**:

  Undergraduate: Thursdays 10-11 LG04, weeks 3-8;
  MSc: Thursdays 11-11 LG03, weeks 4, 5, 7, 8.
- Class tutors:
  - G. Deligiannidis first half, Lawrence Murray second half.
- Hand in solutions by Tuesday, 1pm 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…
- More than $2,000,000$ 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

$$\widehat{I}_n = \frac{1}{n} \sum_{i=0}^{n-1} f\left(\frac{i + 1/2}{n}\right).$$

Riemann Sums

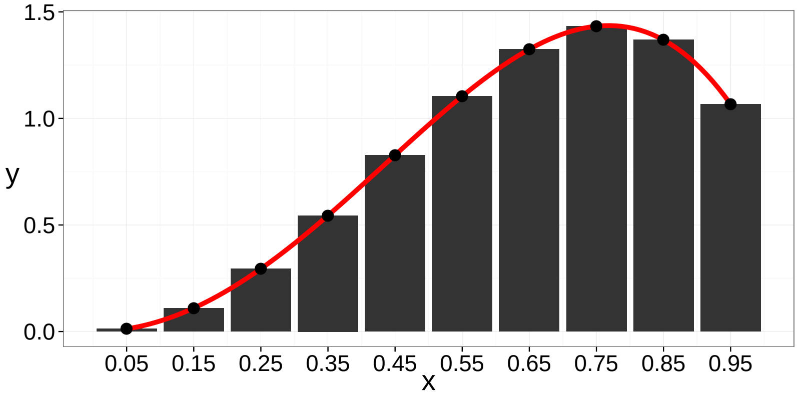

Figure: Riemann sum approximation (black rectangles) of the integral of f (red curve).

Error of naive numerical integration in 1D

- Naively, for a small interval $[a, a+\varepsilon]$ approximate

$$\int_a^{a+\varepsilon} f(x)\mathrm{d}x \approx \varepsilon \times f(a).$$

- Error bounded above by

$$\begin{aligned}\Big|\int_a^{a+\varepsilon} f(x)\mathrm{d}x - \varepsilon \times f(a)\Big| &= \Big|\int_a^{a+\varepsilon} [f(x) - f(a)]\mathrm{d}x\Big| \\ &\le \int_a^{a+\varepsilon}\int_{y=a}^{x} |f'(y)|\mathrm{d}y\,\mathrm{d}x \le \sup_{x\in[0,1]} |f'(x)|\frac{\varepsilon^2}{2}.\end{aligned}$$

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

$$Mn \times \frac{1}{n^2} = \mathcal{O}(1/n).$$

Computing High-Dimensional Integrals

- For $\mathbb{X} = [0,1] \times [0,1]$ using $n = m^2$ evaluations

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

  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 $\mathbb{X} = [0,1]^d$ 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 $f : \mathbb{X} \to \mathbb{R}$, write
$$I = \int_{\mathbb{X}} f(x)\,dx = \int_{\mathbb{X}} \varphi(x)\pi(x)dx.$$
where π is a probability density function on $\mathbb{X}$ and
$$\varphi : x \mapsto f(x)/\pi(x).$$
- Monte Carlo method:
  - sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \pi$,
  - compute
$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i).$$
- Strong law of large numbers: $\widehat{I}_n \to I$ almost surely;
- Central limit theorem: the random approximation error is
$$\mathcal{O}(n^{-1/2})$$
whatever the dimension of the state space $\mathbb{X}$.

Monte Carlo Integration

- In many cases the integral of interest is in the form

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

  for a specific function φ and distribution π.
- The distribution π 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:

$$\begin{aligned}&\text{Monte Carlo method to approximate } \mathbb{E}_{\pi}\left[\varphi(X)\right]\\ &\Leftrightarrow \text{simulation method to sample } \pi\end{aligned}$$

- Thus Monte Carlo sometimes refer to simulation methods.

Ising Model

- Consider a simple 2D-Ising model on a finite lattice $\mathcal{G} = \{1, 2, ..., m\} \times \{1, 2, ..., m\}$ where each site $\sigma = (i, j)$ hosts a particle with a +1 or -1 spin modeled as a r.v. X_σ.
- The distribution of $X = \{X_\sigma\}_{\sigma \in \mathcal{G}}$ on $\{-1, 1\}^{m^2}$ is given by

$$\pi_\beta(x) = \frac{\exp(-\beta U(x))}{Z_\beta}$$

  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_β.
- The dimension is m^2, where m can easily be 10^3.

Ising Model

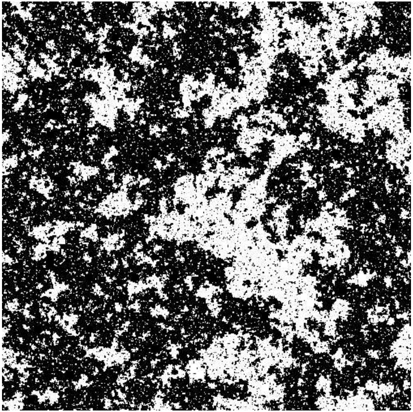

Figure: One draw from the Ising model on a 500×500 lattice.

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

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

Option Pricing

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

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

$$\begin{aligned} S(t+1) &= g(S(t), W(t+1)) \\ &= g(g(S(t-1), W(t)), W(t+1)) \\ &=: g^{t+1}(S(0), W(1), ..., W(t+1)) \end{aligned}$$

  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

$$\begin{aligned}
&\mathbb{E}\left[\max\left(0, S(T) - K\right)\right] \\
&= \int \max\left[0, g^T\left(s(0), w(1), ..., w(T)\right) - K\right] \\
&\quad \times \left\{\prod_{t=1}^{T} p_W\left(w(t)\right)\right\} dw(1) \cdots dw(T),
\end{aligned}$$

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 θ.

- In the frequentist approach θ 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 ϑ 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.

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

- 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*

$$p_{\vartheta|Y}\left(\theta|\, y\right) = \frac{p_Y\left(y;\theta\right) p_\vartheta\left(\theta\right)}{p_Y\left(y\right)}$$

where

$$p_Y\left(y\right) = \int_\Theta p_Y\left(y;\theta\right) p_\vartheta\left(\theta\right) \mathrm{d}\theta$$

is the so-called *marginal likelihood* or *evidence*.

- Point estimates, e.g. posterior mean of ϑ

$$\mathbb{E}\left(\vartheta|y\right) = \int_\Theta \theta \; p_{\vartheta|Y}\left(\theta|\, y\right) \mathrm{d}\theta$$

can be computed.

Bayesian Inference

- Credible intervals: an interval C such that

$$\mathbb{P}\left(\vartheta \in C \middle| y\right) = 1 - \alpha.$$

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

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

- Above predictive density takes into account the *uncertainty about the parameter* θ.
- Compare to simple plug-in rule $p_Y\left(y_{new};\widehat{\theta}\right)$ where $\widehat{\theta}$ is a point estimate of θ (e.g. the MLE).

Bayesian Inference: Gaussian Data

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

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

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

where

$$\omega^2 = \frac{\kappa^2 \sigma^2}{n\kappa^2 + \sigma^2}, \ \nu = \frac{\sigma^2}{n\kappa^2 + \sigma^2}\mu + \frac{n\kappa^2}{n\kappa^2 + \sigma^2}\overline{y}.$$

- Thus $\mathbb{E}(\vartheta | y) = \nu$ and $\mathbb{V}(\vartheta | y) = \omega^2$.

Bayesian Inference: Gaussian Data

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

$$\mathbb{P}(\vartheta \in C \mid y) = 1 - \alpha.$$

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

$$p(y_{n+1} \mid y) = \int_{\Theta} p(y_{n+1} \mid \theta)\, p(\theta \mid y)\, d\theta = \mathcal{N}\left(y_{n+1}; \nu, \omega^2 + \sigma^2\right).$$

- 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

$$\mathbb{P}\left(Y_i = 1 \middle| \theta \right) = \frac{1}{1 + e^{-\theta^T x_i}}$$

- Assign a prior $p(\theta)$ on ϑ. Then Bayesian inference relies on

$$p\left(\theta \middle| y_1, ..., y_n \right) = \frac{p(\theta) \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \middle| \theta \right)}{\mathbb{P}(y_1, ..., y_n)}$$

- If the prior is Gaussian, the posterior is not a standard distribution: $\mathbb{P}(y_1, ..., y_n)$ cannot be computed.

S&P 500 index

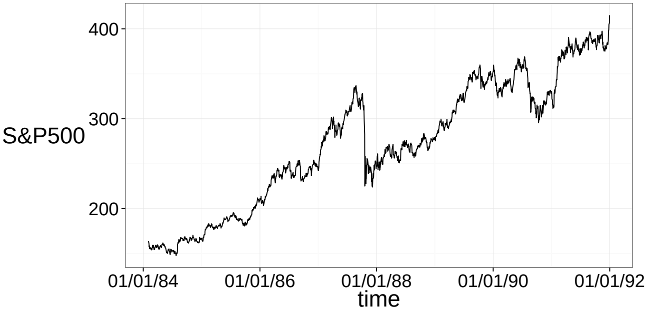

Figure: S&P 500 daily price index (p_t) between 1984 and 1991.

S&P 500 index

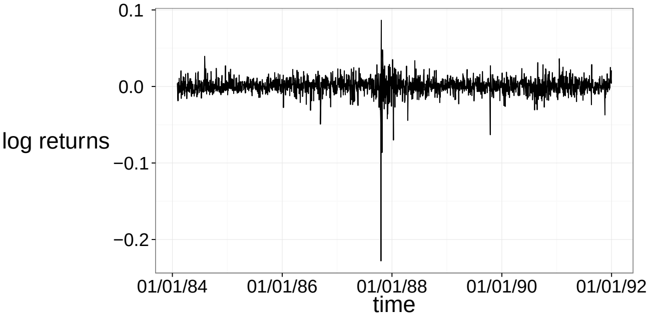

Figure: Daily returns $y_t = \log(p_t / p_{t-1})$ between 1984 and 1991.

Bayesian Inference: Stochastic Volatility Model

- Latent stochastic volatility $(X_t)_{t \geq 1}$ of an asset is modeled through
$$X_t = \varphi X_{t-1} + \sigma V_t,\ Y_t = \beta \exp(X_t) W_t$$
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 (φ, σ, β) 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!