Computational Statistics & Machine Learning

Lecture 2

The Monte Carlo Method

Mark Girolami
mag92@cam.ac.uk

Department of Engineering
University of Cambridge

September 14, 2021

Overview

- Monte Carlo in practice.
- Monte Carlo error.
- Importance Sampling.
- Control Variates.

Vanilla Monte Carlo

- Require methods to simulate random variables exactly from the desired distributions
- There are many suites of comprehensive tools in almost all software libraries for random number generation
- We will work on the basis that these are available to us the practitioners
- We shall later meet the situation where we require samples of random variables where the probability density is not fully known
- For now given a probability density $p(x)$ it is assumed that one is able to both simulate from $p(x)$ and evaluate it

Monte Carlo Error

- The CLT provides

$$\frac{1}{N}\sum_n f(x_n) \sim \mathcal{N}\left(\mathbb{E}\left\{f(X)\right\}, \frac{\sigma_f^2}{N}\right)$$

- The Monte Carlo estimate is unbiased as it converges to $\mathbb{E}\left\{f(X)\right\}$ by Law of Large Numbers (LLN)
- The error bars for an N estimate follow as $\frac{\sigma_f^2}{N}$
- With

$$\sigma_f^2 = \mathbb{E}\{f^2(X)\} - \mathbb{E}\{f(X)\}^2$$

- Note that for any arbitrary dimension of the integral, the error does not depend on the dimensionality of the integral - a major benefit of Monte Carlo

Monte Carlo Error

- Although error reduces as N increases the σ_f^2 term has a significant impact on the error reduction
- Consider the two components of the error - $\mathbb{E}\{f(X)\}^2$ is fixed as a proportion of the squared integral value
- The other component $\mathbb{E}\{f^2(X)\}$ is dependent on the choice of both p and corresponding f
- Different choices of p leads to different level of variance of the estimator
- Consider evaluating $\mathbb{E}\{x^{10}\}$ where expectation is w.r.t. the uniform distribution on the unit line
- Most of the samples would be uninformative, inflating the variance of the estimator

Monte Carlo Error

- Illustrative example
- The lifetime of a system component follows an exponential distribution with a mean of one year i.e.

$$T_{life} \sim p(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- How likely is it that the component survives for a lifetime of 25 (years say) ?
- We need to estimate

$$\begin{aligned} I = P(T_{life} \geq 25) &= \mathbb{E}\{I_{\{T_{life} \geq 25\}}\} \\ &= \int_0^{+\infty} I_{\{x \geq 25\}} p(x) dx \\ &= \int_0^{+\infty} I_{\{x \geq 25\}} e^{-x} dx \end{aligned}$$

Monte Carlo Error

- Monte Carlo estimate is

$$\frac{1}{N}\sum_{n} I_{\{x_n \geq 25\}}$$

  with each $x_n \sim p(x)$
- Try this - what is the problem?
- See Jupyter notebook `Lecture_3.ipynb` for implementation

Importance Sampling

- The problem with the previous example is that the probability density of the random variable gives very small probability to the *important* region of the domain i.e. values of lifetime greater than 25
- N will need to be astronomically large to get anything resembling a reasonably confident Monte Carlo estimate
- The required probability density largely misses the *important* regions of space
- Idea - can one select an alternative density that is aligned with the *important* regions of the space and use that?

Importance Sampling

- We wish to estimate $I = \mathbb{E}_p\{f(X)\} = \int f(x)p(x)dx$
- Another probability density $q(x) \neq 0$ when $p(x) \neq 0$ is selected - e.g. to be better aligned to the important regions - an Importance Distribution
- Then

$$\begin{aligned} I = \mathbb{E}_p\{f(X)\} &= \int f(x)p(x)dx \\ &= \int f(x)p(x)\frac{q(x)}{q(x)}dx \\ &= \int \frac{f(x)p(x)}{q(x)}q(x)dx \\ &= \mathbb{E}_q\left\{\frac{f(X)p(X)}{q(X)}\right\} \end{aligned}$$

- Both the function and integrating density are changed
- Does this provide a solution to the problem?

Importance Sampling

- The original Monte Carlo estimate

$$\frac{1}{N}\sum_n f(x_n)$$

with each $x_n \sim p(x)$ is replaced with the alternative Monte Carlo estimate

$$\frac{1}{N}\sum_n \frac{f(x_n)p(x_n)}{q(x_n)}$$

with each $x_n \sim q(x)$

Importance Sampling

- Sanity check

$$
\begin{aligned}
\mathbb{E}_q\left\{\frac{1}{N}\sum_n \frac{f(x_n)p(x_n)}{q(x_n)}\right\} &= \frac{1}{N}\sum_n \mathbb{E}_q\left\{\frac{f(x_n)p(x_n)}{q(x_n)}\right\} \\
&= \frac{1}{N}\times N \times \mathbb{E}_q\left\{\frac{f(X)p(X)}{q(X)}\right\} \\
&= \mathbb{E}_p\left\{f(X)\right\}
\end{aligned}
$$

- The alternative Monte Carlo estimate is unbiased - as hoped for
- What about the error - the variance of the estimate ?

Importance Sampling

- Monte Carlo Error of Importance Sampler
- Denote $\frac{f(x)p(x)}{q(x)}$ as $\bar{f}(x)$ then

$$\sigma_{\bar{f}}^2 = \mathbb{E}_q\{\bar{f}^2(X)\} - \mathbb{E}_q\{\bar{f}(X)\}^2 = \mathbb{E}_q\{\bar{f}^2(X)\} - I^2$$

- The original estimate has error

$$\sigma_f^2 = \mathbb{E}_p\{f^2(X)\} - \mathbb{E}_p\{f(X)\}^2 = \mathbb{E}_p\{f^2(X)\} - I^2$$

- The difference in errors is

$$\begin{aligned}\sigma_f^2 - \sigma_{\bar{f}}^2 &= \mathbb{E}_p\{f^2(X)\} - \mathbb{E}_q\{\bar{f}^2(X)\} \\ &= \int f^2(x)p(x)dx - \int \frac{f(x)^2 p(x)}{q(x)} p(x)dx\end{aligned}$$

- Clearly (work it out) the biggest decrease in error is achieved when

$$q(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$$

- That is when $\sigma_{\bar{f}}^2 = 0$

Importance Sampling

- Some intuition is helpful at this point
- In regions where the function value $f(x)$ is small or negligible we would want the density to also be small, in other words we do not waste effort in producing samples which do not contribute to the value of the integral
- In regions where the function value $f(x)$ is large the density should also be large so that simulated function evaluations contribute substantively to the value of the integral
- These desiderata are satisfied in the optimal Importance density where $q(x) \propto |f(x)|p(x)$
- Is this useful?

Importance Sampling

- Return now to the illustrative example
- We need to estimate

$$I = P(T_{life} \geq 25) \;=\; \int_0^{+\infty} I_{\{x \geq 25\}} e^{-x} dx$$

- The integrand has value zero below $x = 25$ and one beyond this so moving from $p(x) = I_{x \geq 0} e^{-x}$ to something like $q(x) = \mathcal{N}(\mu, 1)$ where $\mu \geq 25$ will follow the general rule of matching the shape of q with the integrand - what would be the optimal $q(x)$?
- Some schoolboy algebra gives the Importance estimator

$$P(T_{life} \geq 25) = \mathbb{E}_q \left\{ I_{\{x \geq 25\}} \sqrt{2\pi} \exp\left\{ \frac{1}{2}(x - \mu)^2 - x \right\} \right\}$$

with $q(x) = \mathcal{N}(\mu, 1)$ This has -'ve values is it valid?
- How does this compare to the original Monte Carlo estimator ?

Control Variates

- Let's say we want to estimate $\mathbb{E}\{f(X)\}$ and we have some additional information in the form of another function $g(x)$ of which we know the value of $\mathbb{E}\{g(X)\}$
- Denote the Monte Carlo estimates as $\hat{E}_f = N^{-1}\sum_n f(x_n)$ and $\hat{E}_g = N^{-1}\sum_n g(x_n)$
- An unbiased estimate follows as

$$\hat{E}_f^c = \hat{E}_f + c(\hat{E}_g - \mathbb{E}\{g(X)\})$$

- The value of c *controls* the quality of our estimate
- Consider the variance of the new estimate by noting that

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

- Then

$$\mathrm{Var}(\hat{E}_f^c) = \mathrm{Var}(\hat{E}_f) + c^2\mathrm{Var}(\hat{E}_g) + 2c\mathrm{Cov}(\hat{E}_f, \hat{E}_g)$$

Control Variates

- The optimal value of c which will minimise the Monte Carlo error (variance of the control variable estimate) follows using schoolboy calculus to yield

$$c_{opt} = -\frac{\mathrm{Cov}(\hat{E}_f, \hat{E}_g)}{\mathrm{Var}(\hat{E}_g)}$$

- Plugging this value into the expression for the estimator variance gives

$$\mathrm{Var}(E_f^{\hat{c}_{opt}}) = \mathrm{Var}(\hat{E}_f) - \frac{\mathrm{Cov}(\hat{E}_f, \hat{E}_g)^2}{\mathrm{Var}(\hat{E}_g)}$$

- If $\mathrm{Cov}(\hat{E}_f, \hat{E}_g) \neq 0$ the variance of the estimator will be reduced
- The stronger the covariance between the two estimates the greater the reduction in Monte Carlo error

Control Variates

- There is no reason to limit to a single control variate
- Define the following unbiased Monte Carlo estimate

$$\hat{E}_f^c = \hat{E}_f + \sum_{i=1}^{M} c_i(\hat{E}_{g_i} - \mathbb{E}\{g_i(X)\})$$

- Whose variance (convince yourself this is the case) follows as

$$\mathrm{Var}(\hat{E}_f^c) = \mathrm{Var}(\hat{E}_f) + 2\mathbf{c}^\mathsf{T}\mathbf{b} + \mathbf{c}^\mathsf{T}\mathbf{C}\mathbf{c}$$

- The $M \times 1$ vector $\mathbf{c}$ has elements c_i
- The $M \times 1$ vector $\mathbf{b}$ has elements $\mathrm{Cov}(\hat{E}_f, \hat{E}_{g_i})$
- The $M \times M$ matrix $\mathbf{C}$ has elements $\mathrm{Cov}(\hat{E}_{g_i}, \hat{E}_{g_j})$

Control Variates

- Hopefully it is straightforward to see that $\mathbf{c}_{opt} = -\mathbf{C}^{-1}\mathbf{b}$
- This is of course just a least squares linear regression
- The variance achieved is then

$$E_f^{\hat{c}_{opt}} = \hat{E}_f - \mathbf{b}^\mathsf{T}\mathbf{C}^{-1}\mathbf{b}$$

- What does this suggest about the choice of control variates ?
- Implement a simple Monte Carlo estimator with control variates - we do not know the variance or covariance values - these will need a pilot run to obtain estimates