Monte Carlo

Monte Carlo

Monte Carlo is useful to estimate the properties of estimates. Assume that we have a population space with a defined process in it and we want to estimate parameters of the process. We can take samples and estimate those parameters. Further, we construct a sampling distribution. The problem with that is that we don't know the actual values of the parameters. What we can simulate the population process and derive samples from it. In this case we know actual values of estimated parameters, by construction. Hence, construction of sample distribution allows us to conclude if parameters are unbiased, consistent and estimate their variance.

Interview Questions

1. Binomial Tree vs Monte-Carlo (2.55)
    - Monte Carlo has low convergence rate (still not slower then binomial), but path dependence is easy. Early excercise is hard.
    - As a rule of thumb, binomial trie is good for low-dimensional problems involving early excercise, and Monte Carlo for high-dimensional problems involving path-dependence.
2. Longstaff-Schwartz algorithm for pricing an early exercisable options with Monte-Carlo
3. Change of Measure. Give an example when a change of measure is useful in MC
4. Grandma problem with monte carlo
5. 159 q5.20 Given that a stock price at time T is N(100,1), you want to price a digital call struck at 110 by Monte Carlo simulation. What will happen if you do this? Improve the method
6. 190 q 6.4 how would you determine pi by Monte Carlo simulation
7. American Asian Option pricing with monte carlo
8. Central Limit Theorem and types of convergence
9. $E[X/Y]! = E[X]/E[Y]$ - bias exists but becomes negligible as the number of replications increases, and the convergence rate of the estimator is unaffected.

. 3 considerations are particularly important for MC computations: time, bias, variance . Model Discretization error.

Variance Reduction Techniques

1. control variates
    - It exploits information about the errors in estimates of known quantities to reduce the error in an estimate of unknown quantity.
    - The idea is to speedup variance convergence, by doing more computations. This works because MC converges very slowly. To reduce Variance by a factor of 2 we need to perform 4*n computations.

- Let Y_1, Y_2, Y_n be outputs from some replications. Now, for each replication we compute X_i. Assume that $E[X_i]$ is known. Then for any b: $Y_i(b) = Y_i - b(X_i - E[X])$. Then the sample mean $\bar{Y}(b) = \bar{Y} - b(\bar{X} - E[X]) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b(X_i - E[X]))$. As an estimator of $E[Y]$, the control estimato is unbiased because $E[\bar{Y}(b)] = E[\bar{Y} - b(\bar{X} - E[X]) = E[\bar{Y}] = E[Y]$
- The variance: $Var[Y_i(b)] = \sigma_Y^2 - 2b\sigma_X\sigma_Y\rho_{XY} + b^2\sigma_X^2 = \sigma^2(b)$. The optimal coefficient $b^* = \frac{\sigma_Y}{\sigma_X}\rho_{XY}$ Then $\frac{\bar{Y} - b^*(\bar{X} - E[X])]}{Var[\bar{Y}]} = 1 - \rho_{XY}^2$

- A few observations follow from this expression:
  - With the optimal coefficient b^* , the effectiveness of a control variate, as measured by the variance reduction ratio (4.4), is determined by the strength of the correlation between the quantity of interest Y and the control X. The sign of the correlation is irrelevant because it is absorbed in b^*.
  - If the computational effort per replication is roughly the same with and without a control variate, then (4.4) measures the computational speed-up resulting from the use of a control. More precisely, the number of replications of the Yi required to achieve the same variance as n replications of the control variate estimator is $n/(1 - \rho_{XY}^2)$.
  - The variance reduction factor $n/(1 - \rho_{XY}^2)$ increases very sharply as $|\rho_{XY}|$ approaches 1 and, accordingly, it drops off quickly as $|\rho_{XY}|$ decreases away from 1. For example, whereas a correlation of 0.95 produces a ten-fold speedup, a correlation of 0.90 yields only a five-fold speed-up; at $|\rho_{XY}|$ = 0.70 the speed-up drops to about a factor of two. This suggests that a rather high degree of correlation is needed for a control variate to yield substantial benefits.
- In derivative pricing simulations, underlying assets provide a virtually universal source of control variates. If $S(t)$ is an asset price then $e^{-rt}S(t)$ is a martingale Then the control variate estimator is $\frac{1}{n}\sum_{i-1}^{n}(Y_i - b[S_i(T) - e^{rT}S(0)])$ If $Y = e^{-rT}(S(T) - K)^+$
- Any martingale with a known initial value provides a potential control variate precisely becasue its expectation at any future time is its initial value.
- Options or Bonds could potentially become control variates
- Every instrument that serves as a good hedge also seves as a good control variate, if it can be easily priced. Additionaly, dynamic hedge can remove all the variance assuming complete market and continuous trading.

2. Antithetic variates

- The method of antithetic variates attempts to reduce variance by introducing negative dependence between pairs of replications. The most aplicable form is based on the observation that if U is uniformly distributed on $[0, 1]$ then $1 - U$ is too. Hence, if we generate a path using as inputs $U_1, U_2, ...U_n$, we can generate a second path using $1 - U_1, 1 - U_2, ..., 1 - U_n$.
- The variables $U_i and 1 - U_i$ form an antithetic pair in the sence that a large value computed from the first path could be balanced out by the value computed by the antithetic path, thus reducing the variance.
- This observations extend to other distribution through the inverse transform method: $F^{-1}(U)$ and $F^{-1}(-U)$ both have distribution F but are antithetic to each other because F^{-1} is monotone. In particular, if $Z_1, Z_2, ...Z_n i.i.d N(0, 1)$ antithetic variables could be formed by $-Z_1, -Z_2, ... - Z_n$.

3. `stratified sampling`

- Stratified sampling refers broadly to any sampling mechanism that constrains the fraction of observation drawn from specific subsets (or strata) of the sample space. Suppose, more specifically, that our goal is to estimate $E[Y]$ with Y real-valued, and let $A_1, ...A_K$ be disjoint subsets of the real line for which $P(Y \in \cup_i A_i) = 1$. Then $E[Y] = \sum_{i=1}^{K} P(Y \in A_i)E[Y|Y \in A_i] = \sum_{i=1}^{K} p_i E[Y|Y \in A_i]$. In random sampling, we generate independent $Y_i, ..., Y_n$ having the same distributino as Y. The fraction of these samples falling in A_i will not in general equal p_i, though it would approach p_i as the sample size n increased. In stratified sampling, we decide in advance what fraction of the samples should be drawn from each stratum A_i; each observation drawn from A_i is constrined to have the distribution of Y conditional on $Y \in A_i$.

- Stratifying nonuniform distributions. Let F be a CDF on the real line and let

$$F^{-1}(u) = infx : F(x) <= u$$

be its generalised inverse. Given probabilities $p_1, ..., p_K$ summing to 1, define

$$a_0 = -\infty, a_1 = F^{-1}(p_1), a_2 = F^{-1}(p_1 + p_2), ..., a_k = F^{-1}(p_1 + ... + p_K) = F^{-1}(1)$$

Define strata

$$A_i = (a_{i-1}, a_i] fori = 1, ..., K$$

To use the sets $A_1, ...A_K$ for stratified sampling, we need to be able to generate samples of Y conditional on $Y \in A_i$. If $U\ Unif[0,1]$ then $V = a_{i-1} + U(a_i - a_{i-1})$ is uniformly distributed between a_{i-1} and a_i and then $F^{-1}(V)$ has a distribution of Y conditional on $Y \in A_i$

4. Latin hypercube sampling
5. `Moment matching methods`
6. `importance sampling`

- Importance sampling attempts to reduce variance by changing the probability measure from which paths are generated. We change measures to try to give more weight to "important" outcomes thereby increasing sampling efficiency.

Consider the problem of estimating

$$\alpha = E[h(X)] = \int h(x)f(x)dx$$

, where $X \in \Re^d$ is a r.v. with probability density f, and $h : \Re^d \to \Re$. Let g be any other probability density on $\Re^d$ s.t. $f(x) > 0 => g(x) > 0 \forall x$. Then

$$\alpha = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

. This integral can be interpreted as an expectation with respect to the density g:

$$\alpha = \tilde{E}[h(X)\frac{f(x)}{g(X)}]$$

. If X_i are now independent draws from g, the importance sampling estimate associated with g is

$$\tilde{\alpha_g} = \tilde{\alpha_g}(n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)\frac{f(X_i)}{g(X_i)}$$

. The weight $\frac{f(X_i)}{g(X_i)}$ is the likelihood ratio or Radon-Nikodym derivative. If follows that $\tilde{\alpha}_g$ is an unbiased estimator of α

- To compare variances with and without importance sampling it therefore suffices to compare second moments:

$$\tilde{E}[(h(X)\frac{f(X)}{g(X)})^2] = E[(h(X)^2\frac{f(X)}{g(X)})]$$

This could be larger or smaller than the second moment $E[h(X)^2]$ without importance sampling. Successful importance sampling lies in the art of selecting an effective importance sampling density g.

- Nevertheless, this optimal choice of g does provide some useful guidance: in designing an effective importance sampling strategy, we should try to sample in proportion to the product of h and f. In option pricing applications, h is typically a discounted payoff and f is the risk-neutral density of a discrete path of underlying assets. In this case, the "importance" of a path is measured by the product of its discounted payoff and its probability density.

Pricing American Options

- Let $U(t), 0 \leq t \leq T$ be the discounted payoff at time t.
- $T \in [0, T]$ be a class of admissible stopping times.

1. Dynamic programming formulation
   - $\tilde{h}_i$ denotes the payoff function for exercise at t_i, $\tilde{V}_i(x)$ - denotes the value of the option at t_i given $X_i = x$, assuming the option has not previously been exercised. The value is determined recursively as follows:
   $$\tilde{V}_m(x) = \tilde{h}_m(x)$$
   $$\tilde{V}_{i-1}(x) = max\{\tilde{h}_{i-1}(x), E[D_{i-1,i}(X_i)\tilde{V}_i(X_i)|X_{i-1} = x]\}$$
   where $D_{i-1,i}(X_i)$ is a discount factor from t_{i-1} to t_i.
   - Stopping Rules - it is convinient to view the pricing problem through stopping rules and exercise regions.
     - Any stopping time τ determinees a value $V_0^{(\tau)}(X_0) = E[h_\tau(X_\tau)]$,
     - Conversely, any rule assigning a value $\tilde{V}_i(x)$ to each state x and exercise opportunity i, with $\tilde{V}_m = h_m$, determines a stopping rule
     $$\tilde{\tau} = min\{i \in \{1, ..., m\} : h_i(X_i) \geq \tilde{V}_i(X_i)\}$$
     The exercise region determined by $\tilde{V}_i$ at the ith exercise date is the set
     $$\{x : h_i(x) \geq \tilde{V}_i(x)\}$$
     and the continuation region is the complement of this set. The stopping rule $\tilde{\tau}$ can thus also be described as the first time the Markov chain X_i enters an exercise region.
     - Continuation value. The continuation value of an American option with a finite numer of exercise opportunities is the value of holding rather than exercising the option.
     - The continuation value in state x at date t_i is:
     $$C_i(x) = E[V_{i+1}(X_{i+1})|X_i = x], i = 0, ..., m-1$$
     - This also satisfy a dynamic programming recursion: $C_m \equiv 0$ and
     $$C_i(x) = E[max\{h_{i+1}(X_{i+1}), C_{i+1}(i+1)\}|X_i = x], fori = 0, ..., m-1$$
     - The option value is $C_0(X_0)$, the continuation value at time 0.
2. Parametric Approximations
3. Random Tree Methods
4. State-Space Partitioning
5. Stochastic Mesh Methods
6. Regression-Based Methods and Weights
   - Each contunuation value $C_i(x)$ is the regressin of the option value $V_{i+1}(X_{i+1})$ on the current state x. Hence, approximate C_i by a linear combination of known functions of the current state and use regression to estimate the best coefficients in this approximation. This approach is relatively fast and broadly applicable; its accuracy depends on the choice of functions used in the regression.

   - Regression-based methods assume that continuation value has the form
   $$E[V_{i+1}(X_{i+1})|X_i = x] = \sum_{r=1}^{M} \beta_{ir}\phi_r(x) = \beta_i^T \phi(x)$$
   for some basis function ϕ and constants β. Then,
   $$\beta_i = (E[\psi(X_i)\psi(X_i)^T])^{-1} E[\psi(X_i)V_{i+1}(X_{i+1})] \equiv B_\psi^{-1} B_{\psi V}$$

- The coefficients β_{ir} could be estimated from observations of pairs

$$(X_{ij}, V_{i+1}(X_{i+1,j})), j = 1, ..., b$$

, each consisting of the state at time i and the corresponding option value at time i+1.

- The estimate $\hat{\beta}_i$ then defines an estimate $\hat{C}_i(x) = \hat{\beta}_i^T \psi(x)$
- Regression-based Pricing Algorithm (Tsitsiklis and Van Roy)
  - Simulate b independent paths $\{X_{1j}, ..., X_{mj}\}, j = 1, ..., b$ of the Markov chain.
  - At terminal nodes, set $\hat{V}_{mj} = h_m(X_{mj}), j = 1, ..., b$.
  - Apply backward induction: for $i=m-1,...,1,
    - given estimated values $\hat{V}_{i+1,j}, j = 1, ..., b$, use regression as above to calculate $\hat{\beta}_i = \hat{B}_\psi^{-1} \hat{B}_{\psi V}$;
    - set $\hat{V}_{ij} = max\{h_i(X_{ij}, \hat{C}_i(X_{ij}\}, j = 1, ..., b$, where $\hat{C}_i(x) = \hat{\beta}_i^T \psi(x)$
    - Set $\hat{V}_0 = (\hat{V}_{11} + ... + \hat{V}_{ib}/b)$
- Longstaff and Schwartz combine continuation values with their interleaving estimator. Then

$$\hat{V}_{ij} = \begin{cases} h_i(X_{ij}) & h_i(X_{ij}) \geq \hat{C}_i(X_{ij}); \\ \hat{V}_{i+1,j} & h_i(X_{ij} < \hat{C}_i(X_{ij})). \end{cases}$$

- The success of any regression-based approach clearly depends on the choice of basis function. Polynomials are popular choice. Through Taylor expansion any sufficiently smooth value function can be approximated by polynomials.

References:

1. Monte Carlo Methods in Financial Engineering. Paul Glasserman
2. Quant Job Interview Questions and Answers. Mark Joshi, Nick Denson