Variance reduction with stratified sampling for bounded monotone functions

Subhasish Basak, Julien Bect, Emmanuel Vazquez

L2S, CentraleSupélec, Université Paris-Saclay

Context

Variance reduction techniques are used to improve the efficiency of Monte Carlo methods for approximating integrals. In this work we propose an adaptive stratified sampling approach based on theoretical bounds of the strata variances using the monotonicity property of the function of interest.

Problem statement

Consider the problem of estimating

$$\bar{\phi} = \mathbb{E}(\phi(X)) \tag{1}$$

where

- $\phi : \mathbb{R}^+ \to [0, 1]$
- ϕ is continuous
- ϕ is monotonically increasing
- $\phi(0) = 0$ and $\lim_{x\to\infty} \phi(x) = 1$
- The CDF F_X of X is known

Motivation

- In the field of microbiology, Quantitative Risk Assessment (QRA) models are used for estimating the risk of a food borne disease
- QRA models include monotonically increasing functions w.r.t the initial concentration of bacteria
- Given the cost of function evaluation the aim is to reduce the sampling budget compared to the simple Monte Carlo algorithm

How stratification works

Consider the N strata $S_i = [l_i, u_i)$ for $i = 1, 2, \ldots, N$ with:

- stratum probability $\omega_i = P[X \in S_i]$
- stratum variance $\tau_i^2 = \mathrm{Var}(\phi(X)|X \in S_i)$

Then the stratified sampling estimator can be written as

$$\hat{\phi}^{\mathrm{ST}} = \sum_{i=1}^{N} \frac{\omega_i}{n_i} \sum_{j=1}^{n_i} \phi(X_{i,j}) \tag{2}$$

and the variance of this estimator

$$\mathrm{Var}(\hat{\phi}^{\mathrm{ST}}) = \sum_{i=1}^{N} \frac{\omega_i^2 \tau_i^2}{n_i} \tag{3}$$

with

- stratum sample size n_i
- stratum samples $X_{i,j}, j = 1, 2, \ldots, n_i$

The optimal choice for n_i, with total budget $\sum_{i=1}^{N} n_i = n$, obtained by minizing the variance is

$$n_i = \frac{\omega_i \tau_i}{\sum_{i=1}^{N} \omega_i \tau_i} n \tag{4}$$

The variance of the stratified sampling estimator with optimal allocation of sampling budget is

$$\mathrm{Var}(\hat{\phi}^{\mathrm{ST}}_{\mathrm{opt}}) = \frac{1}{n} \left(\sum_{i=1}^{N} \omega_i \tau_i \right)^2 \tag{5}$$

Stratification vs Simple Monte Carlo

- The simple Monte Carlo estimate is a special case of the stratified sampling estimate with $N = 1$
- Both estimates are unbiased
- The variance of the Monte Carlo estimator can be shown to be larger than $\mathrm{Var}(\hat{\phi}^{\mathrm{ST}}_{\mathrm{opt}})$

Conservative approach

- The stratum variances τ_i^2 are unknown
- We propose using upper bounds instead of pilot sample estimates

Popoviciu's inequality provides an upper bound on the variance of bounded random variables

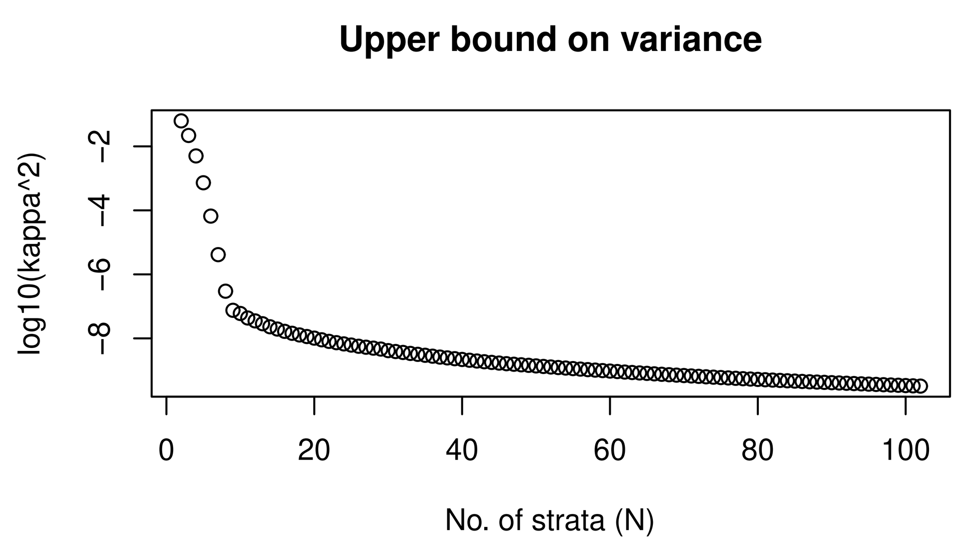
$$\tau_{n,i}^2 \le \frac{1}{4}(\phi(u_i) - \phi(l_i))^2 = \frac{1}{4}\Delta_i^2 \tag{6}$$

- An upper bound of the optimal variance can be written as

$$\mathrm{Var}(\hat{\phi}^{\mathrm{ST}}_{\mathrm{opt}}) \le \frac{1}{4n} \left(\sum_{i=1}^{N} \omega_i \Delta_i \right)^2 = \frac{\kappa^2}{n} \tag{7}$$

- For fixed n, the upper bound $\frac{\kappa^2}{n}$ decreases as the number of strata N increases
- It can be shown that a split in the i-th stratum with any split proportion $0 < \alpha = \frac{\omega_{1,i}}{\omega_i} < 1$ and $0 < \beta = \frac{\Delta_{1,i}}{\Delta_i} < 1$, reduces κ

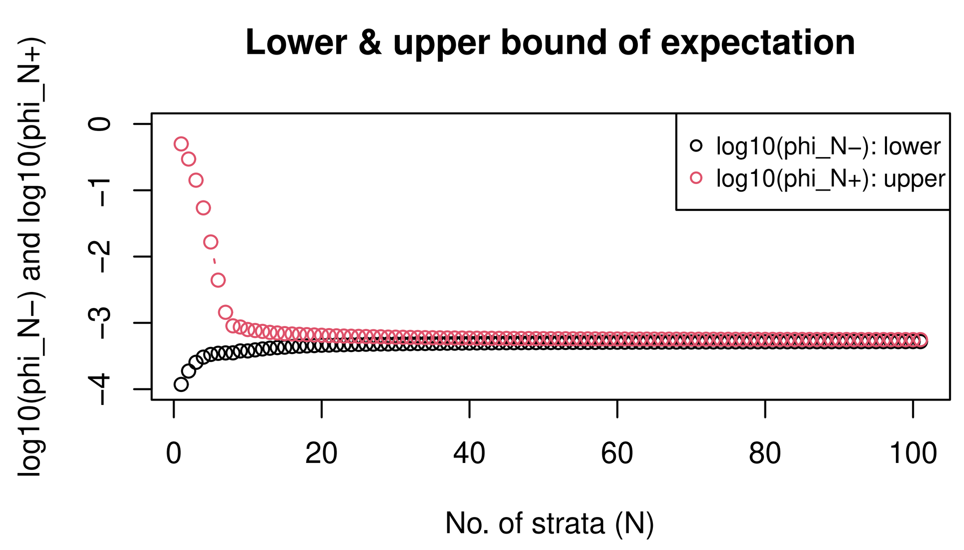
$$\omega_i \Delta_i > \omega_{1,i}\Delta_{1,i} + (\omega_i - \omega_{1,i})(\Delta_i - \Delta_{1,i}) \tag{8}$$



- Bounds on the expectation can be derived using the monotonicity property of ϕ

$$\bar{\phi} \ge \sum_{i=1}^{N} \omega_i \phi_{(l_i)} = \phi_N^-$$
$$\bar{\phi} \le \sum_{i=1}^{N} \omega_i \phi_{(u_i)} = \phi_N^+ \tag{9}$$

- The upper and lower bounds for the expectation converges as number of strata N increases



Adaptive algorithm

- The idea is to split the stratum with maximum $\omega_i \Delta_i$ value (the highest contribution to the variance upper bound)
- A split which divides either ω_i or Δ_i into half, results in reducing the contribution of that particular stratum also by half
- To obtain a split with α or β close to 0.5 we propose the following strategy

$k \leftarrow \arg\max_i(\omega_i \Delta_i)$
if $k = N$ **then**
  $X_{(k)} = 2X_{(k-1)}$
else
  $X_{(k)} = \frac{X_{(k-1)} + X_{(k)}}{2}$

- For a fixed number of strata N the **sampling budget** n can be obtained by fixing the upper bound of coefficient of variation by δ

$$\mathrm{CV}(\hat{\phi}^{\mathrm{ST}}) = \frac{\sqrt{\mathrm{Var}(\hat{\phi}^{\mathrm{ST}})}}{\bar{\phi}} \le \frac{\kappa}{\sqrt{n}\phi_N^-} \le \delta \tag{10}$$
$$\implies n = \left\lceil \frac{\kappa^2}{(\phi_N^-)^2 \delta^2} \right\rceil$$

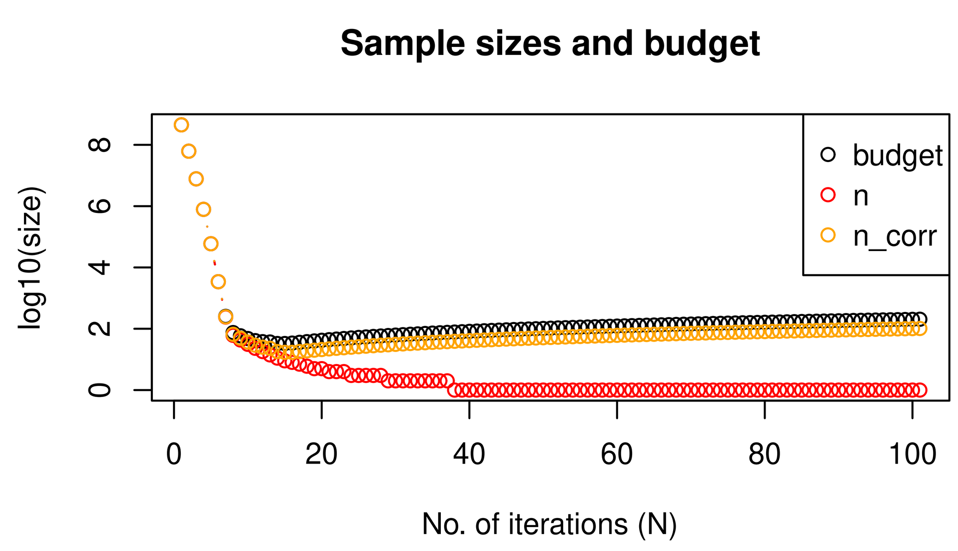
- The optimal strata sizes given by (4) (by substituting Δ_i) might result in non integer values
- To ensure each stratum has at least one sample, we recompute the budget as n^{corr} by taking the ceiling value

$$n^{\mathrm{corr}} = \sum_{i=1}^{N} \tilde{n}_i = \sum_{i=1}^{N} \lceil n_i \rceil \ge n \tag{11}$$

- The upper bound for variance with the corrected sample size is smaller

$$\sum_{i=1}^{N} \frac{\omega_i^2 \Delta_i^2}{n_i} \ge \sum_{i=1}^{N} \frac{\omega_i^2 \Delta_i^2}{\tilde{n}_i} \tag{12}$$

- The actual **budget** of the algorithm is n^{corr}+ the additional evaluations made in each strata for splitting



Stopping rule: The algorithm stops splitting the strata as the budget starts increasing

Experimental results

- The proposed method is implemented and compared to simple Monte Carlo for estimating the risk in a QRA model
- A simple Monte Carlo approach requires a budget of 2503 samples to achieve a 10% CV
- The proposed stratified sampling approach requires 33 samples to achieve the same CV