Integration of bounded monotonic functions:

Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods

Subhasish Basak[1,2], Julien Bect[2], Emmanuel Vazquez[2]

JdS 2022, Lyon, 15 Juin

[1]ANSES, Maisons-Alfort, France

[2]Univ. Paris-Saclay, CNRS, CentraleSupeléc, L2S, Gif-sur-yvette, France

Problem statement

We are interested in approximating $\mathbb{E}(g(Y)) = \int g(y)\, \mathrm{P}_Y(\mathrm{d}y)$

- $g : \mathbb{R} \to \mathbb{R}$ is monotonic and bounded
- P_Y is known, CDF F_Y of Y is continuous
- After scaling, the problem reduces to estimating

$$S(f) = \int_0^1 f(x)\mathrm{d}x\,,$$

  where $f : [0,1] \to [0,1]$ is a non-decreasing function

- We consider a fixed sample-size setting where $S(f)$ is estimated from n evaluations of f

Example: Quantitative Risk Assessment (QRA)

- Consider the following example in food safety
  - Y: dose of pathogen STEC (Shiga Toxin producing Escherichia coli)
  - $g(Y)$: conditional risk of food-borne disease (Haemolytic Uremic Syndrome)
  - g is an increasing function (complex numerical model)
- Goal: estimate $\mathbb{E}(g(Y))$

Contents

1 Literature review

- Two classes of methods (sequential or nonsequential):
  - **deterministic**
  - **randomized**
- Kiefer (1957) studied deterministic methods
  - Trapezoidal rule is the optimal worst-case deterministic algorithm

– It attains the lower bound for the maximal error: $\frac{1}{2\cdot(n+1)}$

- Novak (1992) studied unbiased randomized nonsequential methods
  - Unbiasedness comes at the price of a slight loss of performance w.r.t. the trapezoidal rule
- Novak (1992) also studied randomized sequential methods
  - He proposed a rate-optimal algorithm ($n^{-3/2}$ for the L_1 error)

- We focus on nonsequential unbiased Monte Carlo methods
- We provide a lower bound for the maximal Lp error
- We study methods based on uniform i.i.d sampling and statified sampling
- These unbiased methods can be used as building blocks for sequential methods

2 Non sequential randomized methods

- A nonsequential randomized method evaluates the function at n random points $X_1, \dots, X_n$ in $[0,1]$ and approximates the integral $S(f)$ using an estimator

$$\widehat{S}_n(f) = \varphi\left(X_1,\, f(X_1),\, \dots,\, X_n,\, f(X_n)\right) \tag{1}$$

- The worst-case L^p error of such a method over the class F is

$$e_p(\widehat{S}_n) \;=\; \sup_{f \in F} \mathbb{E}\left(\left|S(f) - \widehat{S}_n(f)\right|^p\right)^{1/p} \tag{2}$$

F is the class of all non-decreasing functions $f : [0,1] \to [0,1]$

3 A lower bound for the L_p error

- We show that (extending Novak's result)

Theorem

For any nonsequential randomized method with sample size n,

$$e_p(\widehat{S}_n) \;\geq\; \left(\frac{1}{2}\right)^{2+1/p} \frac{1}{n}\,.$$

Corollary

$$\sup_{f \in F} \operatorname{Var}\left(\widehat{S}_n(f)\right) \;\geq\; \frac{1}{32n^2}.$$

4 Uniform i.i.d sampling

- We study unbiased nonsequential methods based on uniform i.i.d sampling
- $\widehat{S}_n(f)$ uses uniform i.i.d samples $\{X_1, X_2, \cdots, X_n\}$ from $[0, 1]$
- We study two unbiased estimators
  - Simple Monte Carlo estimator
  - Control variate estimator

- Simple Monte Carlo estimator $\widehat{S}_n^{\mathrm{MC}}(f)$
  - The estimator is defined as

$$\widehat{S}_n^{\mathrm{MC}} = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad X_i \overset{iid}{\sim} U_{[0,1]} \tag{3}$$

  - Using Popoviciu's inequality:

$$\left(e_2\left(\widehat{S}_n^{\mathrm{MC}}\right)\right)^2 = \max_{f\in F} \mathrm{Var}\left(\widehat{S}_n^{\mathrm{MC}}(f)\right) = \frac{1}{4n}.$$

  - The maximal error is attained when f is a unit step function jumping at $x_0 = 1/2$

- Control variate estimator $\widehat{S}_n^{\mathrm{cv}}(f)$
  - We consider the control variate $\tilde{f}(X_i) = X_i$ and the estimator

$$\widehat{S}_n^{\mathrm{cv}}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i) - \tilde{f}(X_i)\Big) + \frac{1}{2}. \qquad (4)$$

  - Then,

$$\left(e_2\left(\widehat{S}_n^{\mathrm{cv}}\right)\right)^2 \;=\; \max_{f\in F} \operatorname{Var}\left(\widehat{S}_n^{\mathrm{cv}}(f)\right) \;=\; \frac{1}{12n}.$$

  - The maximal error is attained for any unit step function

5 Stratified sampling

- Unbiased randomized nonsequential methods based on stratified sampling
- $[0,1]$ is divided into K strata, with the k-th stratum denoted by $I_k = [x_{k-1}, x_k]$, $0 = x_0 < x_1 < \ldots < x_{K-1} < x_K = 1$
  - stratum size: $w_k = |x_k - x_{k-1}|$
  - stratum budget: n_k
- $\widehat{S}_n(f)$ uses n_k uniformly distributed samples from I_k

- Stratified sampling estimator $\widehat{S}_n^{\text{str}}(f)$
  - We consider K strata with sizes w_k and budgets n_k

$$\widehat{S}_n^{\text{str}}(f) = \sum_{k=1}^{K} w_k \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} f\left(X_{k,i}\right), \tag{5}$$

  - Then,

$$\left(e_2\left(\widehat{S}_n^{\text{str}}\right)\right)^2 = \max_{f \in F} \operatorname{Var}\left(\widehat{S}_n^{\text{str}}(f)\right) = \frac{1}{4} \max_k \frac{w_k^2}{n_k},$$

  - The minimal value of the maximal error is $\frac{1}{4 \cdot n^2}$, attained with $K = \sum_k n_k$ and $w_k = n^{-1}$.

6 Discussion

Maximal error of nonsequential randomized methods

lower bound	$\widehat{S}_n^{\mathrm{str}}(f)$	$\widehat{S}_n^{\mathrm{cv}}(f)$	$\widehat{S}_n^{\mathrm{MC}}(f)$
$1/32 \cdot n^2$	$1/4 \cdot n^2$	$1/12 \cdot n$	$1/4 \cdot n$

- The best-known variance upper bound
  - $n \leq 2$: Control variate estimator $\widehat{S}_n^{\mathrm{cv}}(f)$
  - $n \geq 3$: Stratified sampling estimator $\widehat{S}_n^{\mathrm{str}}(f)$

(The squared error for the deterministic biased trapezoidal rule is $\frac{1}{4(n+1)^2}$
$\rightarrow$ outperforms randomized unbiased methods.)

Ratio of variance upper bound to lower bound

$n=1$	$n=2$	$n \geq 3$
2.67	5.33	8

- We do not know at the moment if these results are optimal in the class of unbiased nonsequential methods
- The lower bound of variance is not known to be sharp
- Future work:
  - Multivariate case with partial monotonicity
  - Constructing sequential randomized methods

References

[1] J. Kiefer. Optimum sequential search and approximation methods under minimum regularity assumptions. Journal of the Society for Industrial and Applied Mathematics, 5(3):105–136, 1957.

[2] M. D. McKay, R. J. Beckman, and C. W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2):239–245, 1979.

[3] E. Novak. Quadrature formulas for monotone functions. Proceedings of the American Mathematical Society, 115(1):59–68, 1992.

[4] T. Popoviciu. Sur l approximation des fonctions convexes d ordre superieur. Mathematica (Cluj), 10:49–54, 1935.

This work is part of the ArtiSaneFood project (grant number : ANR-18-PRIM-0015) which is part of the PRIMA program supported by the European Union.

Integration of bounded monotonic functions:

Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods

Subhasish Basak[1,2], Julien Bect[2], Emmanuel Vazquez[2]

JdS 2022, Lyon, 15 Juin

[1]ANSES, Maisons-Alfort, France

[2]Univ. Paris-Saclay, CNRS, CentraleSupeléc, L2S, Gif-sur-yvette, France