INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

Working Paper

IIMK/WPS/337/ITS/2019/01

MARCH 2019

ANOVA with two timescale stochastic approximation for estimating Variance of Conditional Expectation

Mohammed Shahid Abdulla[1]
L Ramprasath[2]

[1]Associate Professor, Information Technology and Systems, Indian Institute of Management, Kozhikode, IIMK Campus PO, Kunnamangalam, Kozhikode, Kerala 673570, India; Email: shahid@iimk.ac.in, Phone Number (+91) 495 – 2809254

[2]Associate Professor, Finance, Accounting and Control , Indian Institute of Management, Kozhikode, IIMK Campus PO, Kunnamangalam, Kozhikode, Kerala 673570, India; Email: lrprasath@iimk.ac.in, Phone Number (+91) 495 – 2809248

ANOVA with two timescale stochastic approximation for estimating Variance of Conditional Expectation

Mohammed Shahid Abdulla† & L Ramprasath*
† IT and Systems Area, IIM Kozhikode
* Finance, Accounting and Control Area, IIM Kozhikode

Abstract

The ANOVA method is of value to detect if a population, consisting of labelled sub-populations, has any statistically significant support for considering such labels as valid. In classical ANOVA, the effect of a variable in each sub-population is treated as a Conditional Expectation (CE), and the variance of such CE among the sub-populations has a bearing on whether the null hypothesis can be rejected or not. ANOVA formulae can therefore be used to estimate the Variance of CE (Var-of-CE) itself, and a fairly recent publication has proposed a method wherein a fixed number of samples in each sub-population is used to estimate Var-of-CE. This method assumes repeated sampling of both sub-populations and samples within them, and have designed provably unbiased estimators of Var-of-CE, with one of these being approximately minimum variance under some conditions. Combined with another more recent method, such methods have disadvantages, such as requiring a pilot simulation, or suffering an empirically-observed Root Mean Squared Error (RMSE) that is unfavourable. The work explained here proposes an ANOVA estimator for Var-of-CE that requires an increasing number of samples from each subpopulation. Yet, the estimator reduces the empirically-observed MSE in Var-of-CE estimate in 3 benchmark experiments from the literature.

I. Introduction

The Analysis of Variance (ANOVA) technique is used extensively in statistical methods to understand whether effect τ_k of an alternate hypothesis has a dominating impact on error $\epsilon_{k,j}$, which is the noise in observation j that belong to a sub-population k. The samples drawn are $X_{k,j}$ where each $X_{k,j} = \mu+\tau_k+\epsilon_{k,j}$, in which μ is expectation of $X_{k,j}$ over all effects k (representing the outer loop of the simulation) and inner loop samples j. In particular, $\mu = \lim_{K\to\infty,n\to\infty} \frac{1}{nK}\sum_{k=1}^{K}\sum_{j=1}^{n} X_{k,j}$. Note that we have assumed n corresponding to an outer loop iteration k to be fixed, but in general it can be an integer n_k dependant on k. Thus K sub-populations are considered in the outer loop of the simulation, while within each sub-population k, n_k samples are considered in the inner loop.

The standard formulas used in ANOVA to estimate Var-of-CE, the quantity below with notation $\hat{\sigma}^2_\tau$, are written as follows:

$$
\begin{aligned}
SS_\epsilon &= \sum_{k=1}^{K}\sum_{j=1}^{n_k}(X_{k,j}-\bar{X}_k)^2 \text{ where } \bar{X}_k := \frac{1}{n_k}\sum_{j=1}^{n_k} X_{k,j} \\
SS_\tau &= \sum_{k=1}^{K} n_k\cdot(\bar{X}_k-\bar{\bar{X}})^2, \text{ where } \bar{\bar{X}} := \frac{1}{C}\sum_{k=1}^{K} n_k\cdot\bar{X}_k, \text{ and } C := \sum_{k=1}^{K} n_k \\
\hat{\sigma}^2_\epsilon &= \frac{SS_\epsilon}{C-K} \qquad (1)\\
\hat{\sigma}^2_\tau &= \frac{SS_\tau-(K-1)\cdot\hat{\sigma}^2_\epsilon}{C-\frac{\sum_{k=1}^{K} n_k^2}{C}} \qquad (2)
\end{aligned}
$$

We have used the notation in [1, (6)-(8)], where a simple derivation of the above formulas is also given.

We consider situations where $K\to\infty$. Note that classical single-factor ANOVA considers finite K and $F = \frac{\left(\frac{SS_\tau}{K-1}\right)}{\hat{\sigma}^2_\epsilon}$ is used as the $F-$statistic with $(K-1, C-K)$ degrees of freedom. If this $F \geq F_\alpha$, where α is a statistical signficance level that depends on degrees of freedom, then the null hypothesis is rejected. Note the requirement that it is sufficient for an unbiased estimator that 1. $K \to \infty$ as $C \to \infty$ - so that estimator $\hat{\sigma}^2_\tau$ has lower variance - and 2. as $K \to \infty$ (therefore $k \to \infty$ also), we require $n_k \to \infty$ to result in lower bias. This implies that a static sampling budget C, which assures nearly unbiased and minimal-variance behaviour, could be such that $K >> 0$ and $n_k = N >> 0$. We utilise this scheme to structure K, $\{n_k\}_{k=1}^K$ to satisfy the above conditions such that 1. $\sum_{k=1}^{K} n_k <= C$ while $\sum_{k=1}^{K+1} n_k <= C$, and 2. $n_k = k^\alpha$, $\alpha > 0$, respectively. The value of α will be further filtered to also satisfy the conditions of two-timescale stochastic approximation.

A. Survey of Literature

Recent work [1] has proposed a one-and-half level nested simulation where a pilot experiment, costing about 20% of sampling budget C, calculates an approximation to the optimal n^*, with $n_k = n^*$ for all k. Here n^* is inner loop size that results in a minimum-variance Var-of-CE estimator $\hat{\sigma}^2_M$, under the conditions that i. $n_k = n$, i.e. n_k is a fixed integer n for all $k \leq K$, such that ii. number of subpopulations $K \to \infty$. More recently, [2], proposed an easier estimator that required only a one level simulation such that $n_k = 2$, $\forall k \leq K$. The advantage with the algorithm in [2] is that it doesn't require a pilot simulation unlike [1]. After establishing that the algorithm is unbiased, [2] test their work on 3 experiments where closed-form value of σ^2_τ is known and thus a diminishing root mean square error (RMSE) is observed against sampling budget C. In contrast, [1] use a Delta Hedging example from finance where variance of estimator $\hat{\sigma}^2_\tau$ in different experiments is recorded, to indicate a low variance when $n_k = n^*$, and higher variances when $n_k = n \neq n^*$, for different C. Notice that RMSE in [1, (10)] is $O(\frac{1}{\sqrt{C}})$ asymptotically despite $n_k = n^*$, i.e. samples drawn from sub-populations k being bounded in number. Notice also that this claim holds true for ANOVA-based Var-of-CE estimator (1)-(2) above.

Note that rate of convergence being $O(\frac{1}{\sqrt{C}})$ would also depend on the method of apportioning K and n_k. For example, one such scheme could be $K = \sqrt{C}$, while $n_k = n$, with $n = \sqrt{C}$. Such a scheme of apportioning, since $n_k = n$, $\forall k$, would nevertheless have the desirable property of RMSE converging at rate $O(\frac{1}{\sqrt{C}})$. Also note in this scheme that as $C \to \infty$, we have $K \to \infty$ and $n \to \infty$, for better properties of the ANOVA Var-of-CE estimator. However, since n_k, $\forall k$, must be calculated upfront, the sampling budget C must also be declared apriori and is therefore not sequential in nature. Separately, experimental performance indicates $O(\frac{1}{\sqrt{C}})$ or better RMSE convergence for ANOVA and 2TS-ANOVA, since these have a lower RMSE than the estimator in [2], as seen below in the results section.

II. Proposed Algorithm: 2TS-ANOVA

The proposed algorithm in this work is to calculate an additional term $\tilde{\bar{\bar{X}}}_k$ which is a critic, updated over n_k samples, for an imaginary actor recursion. The actor role is also played by $\tilde{\bar{\bar{X}}}_k$, with the constraint that its value is evaluated only at the $k-$th outer loop instance. The requirement in 2TS algorithms is that changes in actor parameter converge to 0 as $k \to \infty$, note that $\tilde{\bar{\bar{X}}}_k \to \bar{\bar{X}}$ as $k \to \infty$ irrespective of any structure on n_k. Therefore the claim is that $\tilde{\bar{\bar{X}}}_k$ converges to $\bar{\bar{X}}$ at the rate $\frac{1}{k}$ and a critic recursion may be designed around this. This critic recursion would also be the calculation of $\tilde{\bar{\bar{X}}}_k$, however at a more granular number of samples, $N_k = \sum_{s=1}^{k} n_s$, with n_s suitably chosen.

The principles of two-timescale actor-critic stochastic approximation algorithms were first proposed in [3]. Note that in [3], it is required to have separating updating stepsizes for both actor and critic algorithms. The specific variant of two-timescale algorithm used here is referred to as a Type-1 algorithm described in [4]. To give an illustration of what an updating stepsize is, assume that $n_k = n$, $\forall k$, and note the recursion $\bar{\bar{X}}_k := \bar{\bar{X}}_{k-1} + \frac{1}{k} \cdot (\bar{X}_k - \bar{\bar{X}}_{k-1})$ for $\bar{\bar{X}}_k \to \bar{\bar{X}}$ used above. In this illustration, $\frac{1}{k}$ is the updating stepsize and also suits the context where $\bar{\bar{X}}$ is calculated as an average of $\bar{X}_k$.

The formulas used in 2TS-ANOVA are proposed as follows:

$$
\begin{aligned}
\tilde{SS}_{\epsilon} &= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(X_{k,j}-\tilde{\bar{X}}_k\right)^2, \text{ where } \tilde{\bar{X}}_k := \frac{1}{N_k}\sum_{m=1}^{k} n_m \cdot \bar{X}_m, \text{ and } N_k := \sum_{m=1}^{k} n_m \\
\tilde{SS}_{\tau} &= \sum_{k=1}^{K} n_k \cdot \left(\bar{X}_k-\tilde{\bar{X}}_k\right)^2 \\
\begin{pmatrix}\hat{\sigma}_{\tau}^2\\ \hat{\sigma}_{\epsilon}^2\end{pmatrix} &= \begin{pmatrix} \sum_{k=1}^{K} n_k \cdot \frac{\sum_{m=1}^{k-1} n_m^2+(N_k-n_k)^2}{N_k^2} & \sum_{k=1}^{K}(1-\frac{n_k}{N_k})^2+\frac{n_k\cdot N_{k-1}}{N_k^2} \\ \sum_{k=1}^{K} n_k \cdot \frac{\sum_{m=1}^{k-1} n_m^2+(N_k-n_k)^2}{N_k^2} & \sum_{k=1}^{K} n_k \cdot \frac{N_k(N_k-1)}{N_k^2}. \end{pmatrix}^{-1} \cdot \begin{pmatrix}\tilde{SS}_{\tau}\\ \tilde{SS}_{\epsilon}\end{pmatrix}
\end{aligned}
$$

The method of calculating coefficients is similar to the treatment in [1, (6)-(8)].

We denote these calculations below:

$$
\begin{aligned}
\tilde{SS}_\epsilon &= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(X_{k,j}-\frac{1}{N_k}\sum_{s=1}^{N_k}X_{s_1,s_2}\right)^2,\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(\mu+\tau_k+\epsilon_{k,j}-\frac{1}{N_k}\sum_{s=1}^{N_k}(\mu+\tau_{s_1}+\epsilon_{s_1,s_2})\right)^2,\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(\tau_k+\epsilon_{k,j}-\frac{1}{N_k}\sum_{s=1}^{N_k}(\tau_{s_1}+\epsilon_{s_1,s_2})\right)^2,\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(\tau_k-\frac{1}{N_k}\sum_{s=1}^{k}n_s\tau_s+\epsilon_{k,j}-\frac{1}{N_k}\sum_{s=1}^{N_k}\epsilon_{s_1,s_2}\right)^2,\\
E(\tilde{SS}_\epsilon) &= \sum_{k=1}^{K}\sum_{j=1}^{n_k}E\left(\tau_k-\frac{1}{N_k}\sum_{s=1}^{k}n_s\tau_s\right)^2+E\left(\epsilon_{k,j}-\frac{1}{N_k}\sum_{s=1}^{N_k}\epsilon_{s_1,s_2}\right)^2\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left((1-\frac{n_k}{N_k})\tau_k-\frac{1}{N_k}\sum_{s=1}^{k-1}n_s\tau_s\right)^2+\left((1-\frac{1}{N_k})\epsilon_{k,j}-\frac{1}{N_k}\sum_{s=1,(s_1,s_2)\neq(k,j)}^{N_k}\epsilon_{s_1,s_2}\right)^2\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left((1-\frac{n_k}{N_k})^2-\frac{1}{N_k^2}\sum_{s=1}^{k-1}n_s^2\right)\sigma_M^2+\left((1-\frac{1}{N_k})^2+\frac{1}{N_k^2}(N_k-1)\right)\sigma_\epsilon^2\\
&= \sum_{k=1}^{K}\sum_{j=1}^{n_k}\frac{\sum_{s=1}^{k-1}n_s^2+(N_k-n_k)^2}{N_k^2}\sigma_M^2+\sum_{k=1}^{K}\sum_{j=1}^{n_k}\frac{(N_k-1)^2+(N_k-1)}{N_k^2}\sigma_\epsilon^2\\
&= \sum_{k=1}^{K}n_k\frac{\sum_{s=1}^{k-1}n_s^2+(N_k-n_k)^2}{N_k^2}\sigma_M^2+\sum_{k=1}^{K}n_k\frac{N_k(N_k-1)}{N_k^2}\sigma_\epsilon^2\\
\tilde{SS}_\tau &= \sum_{k=1}^{K}n_k(\bar{X}_k-\tilde{\bar{X}}_k)^2,\\
&= \sum_{k=1}^{K}n_k\left(\frac{1}{n_k}\sum_{j=1}^{n_k}X_{k,j}-\frac{1}{N_k}\sum_{j=1}^{N_k}X_{j_1,j_2}\right)^2,\\
E(\tilde{SS}_\tau) &= \sum_{k=1}^{K}n_k\cdot E\left((\tau_k-\frac{1}{N_k}\sum_{j=1}^{k}n_j\tau_j)+(\bar{\epsilon}_k-\frac{1}{N_k}\sum_{j=1}^{k}n_j\bar{\epsilon}_j)\right)^2,\\
&= \sum_{k=1}^{K}n_k\cdot\left(\left((1-\frac{n_k}{N_k})^2+\frac{1}{N_k^2}\sum_{j=1}^{k-1}n_j^2\right)\cdot\sigma_M^2+\left((1-\frac{n_k}{N_k})^2\frac{1}{n_k}+\frac{1}{N_k^2}\sum_{j=1}^{k-1}n_j\right)\cdot\sigma_\epsilon^2\right),\\
&= \sum_{k=1}^{K}n_k\cdot\left((1-\frac{n_k}{N_k})^2+\frac{1}{N_k^2}\sum_{j=1}^{k-1}n_j^2\right)\cdot\sigma_M^2+\sum_{k=1}^{K}n_k\cdot\left((1-\frac{n_k}{N_k})^2\frac{1}{n_k}+\frac{1}{N_k^2}\sum_{j=1}^{k-1}n_j\right)\cdot\sigma_\epsilon^2,\\
&= \sum_{k=1}^{K}n_k\cdot\frac{((N_k-n_k)^2+\sum_{j=1}^{k-1}n_j^2)}{N_k^2}\cdot\sigma_M^2+\sum_{k=1}^{K}\left((1-\frac{n_k}{N_k})^2+\frac{n_k\cdot N_{k-1}}{N_k^2}\right)\cdot\sigma_\epsilon^2.
\end{aligned}
$$

The method of calculating coefficients is similar to the treatment in [1, (6)-(8)].

III. Results

Consider Experiment 3 of [2] as the first example, and $n_k=\lceil k^{0.51}\rceil$, where 0.51 is used such that $n_k\to\infty$ as $k\to\infty$ (a requirement in [3]). The outcome in Experiments 1-2 of [2] also indicate advantage for the

2TS-ANOVA algorithm, with performance of algorithm in [2] in Experiment 3 also plotted for reference below. The second graph below is to compare the variance of ANOVA and ANOVA-2TS in the Delta Hedging setting of [1]. The measured variance is equivalent for lower sampling budgets - with a slight edge for ANOVA-2TS - but becomes indistinguishable later.

We describe the 3 experiments from [2] as follows: in the first experiment, a random variable Y_k is sampled from the distribution $\beta(4,4)$, then samples $\{X_{k,j}\}_{j=1}^{n_k}$ are sampled from $N(Y_k, \sqrt{0.5})$. Note that for the method of [2], $n_k = 2$, $\forall k$. For the second experiment, Y_k is sampled as in the first experiment, but samples $\{X_{k,j}\}_{j=1}^{n_k}$ are drawn from $N(Y_k, Y_k)$. In the third example, the inner-loop samples $\{X_{k,j}\}_{j=1}^{n_k}$ are drawn from exponential distribution as $\mathrm{EXP}(\frac{1}{Y_k+1})$. Experimental results of all 3 experiments are included here, over 10000 simulations each, and are compared with [2] algorithm.

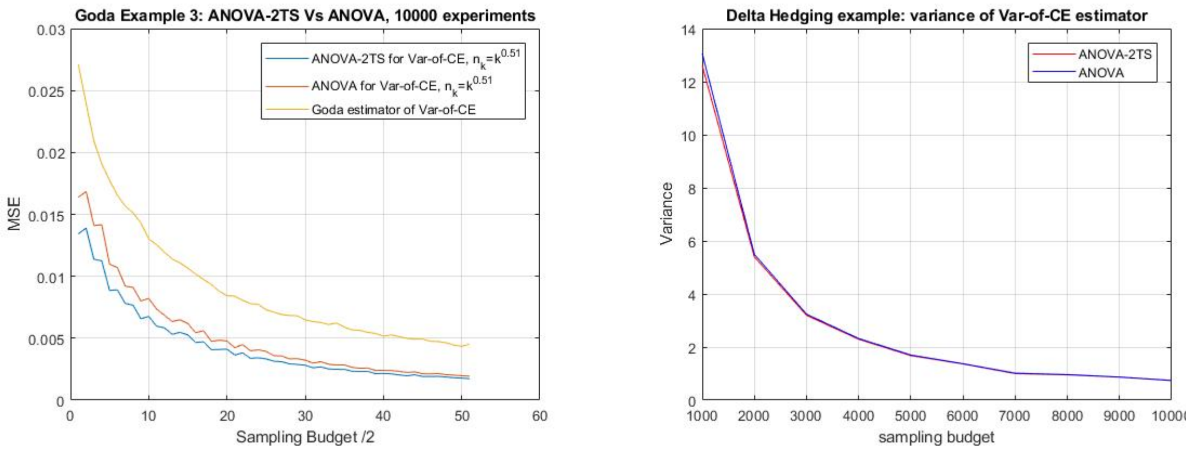

Fig. 1. Performance of ANOVA Vs 2TS-ANOVA in Example 3 of [2] and [1]

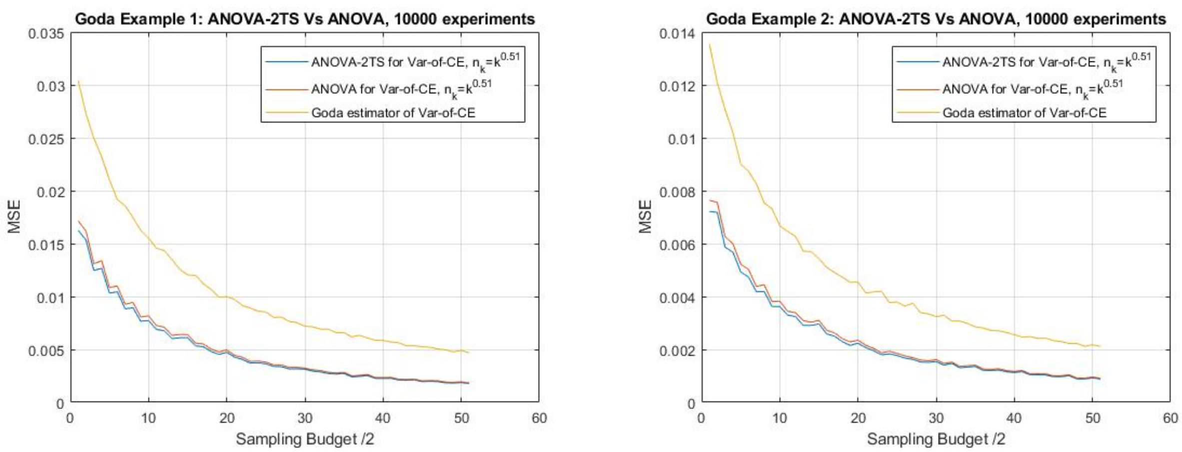

Fig. 2. Performance of ANOVA Vs 2TS-ANOVA in Examples 1 and 2 of [2]

References

[1] Y. Sun, D. W. Apley, and J. Staum, "Efficient nested simulation for estimating the variance of a conditional expectation," Operations Research, vol. 59, no. 4, pp. 998–1007, 2011.

[2] T. Goda, "Computing the variance of a conditional expectation via non-nested monte carlo," Operations Research Letters, vol. 45, pp. 63–67, 2017.

[3] V. S. Borkar, "Stochastic approximation with two time scales," Systems and Control Letters, vol. 29, no. 5, pp. 291–294, 1997.

[4] S. Bhatnagar, M. C. Fu, S. I. Marcus, and I.-J. Wang, "Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences," ACM Transactions on Modeling and Computer Simulation, vol. 13, no. 2, pp. 180–209, 2003.

Research Office

Indian Institute of Management Kozhikode

IIMK Campus P. O.,

Kozhikode, Kerala, India,

PIN - 673 570

Phone: +91-495-2809237/ 238

Email: research@iimk.ac.in

Web: https://iimk.ac.in/faculty/publicationmenu.php