# Diagnosing Learning Algorithms with Super-optimal Recursive Estimators (No. 704)

Man Fung Leung, Kin Wai Chan

Department of Statistics, The Chinese University of Hong Kong

ICSA 2020 Applied Statistics Symposium

# Elevator Speech

# Introduction

Consider the estimation of sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ where:

1. the data $X_i$ can be serially dependent;
2. the data $X_i$ arrives sequentially;
3. the sample size $n$ is not known *a priori*.

Two ways to compute $\bar{X}_n$:

1. (Non-recursive) calculate $(X_1 + X_2 + \cdots + X_n)/n$;
    1. $O(n)$-time update: need to add up $n$ elements.
    2. $O(n)$-space update: need to remember $n$ elements.
2. (Recursive) calculate $\{(n-1)\bar{X}_{n-1} + X_n\}/n$.
    1. $O(1)$-time update: need to add up 2 elements only.
    2. $O(1)$-space update: need to remember 2 elements only.

This setting appears frequently with the use of learning algorithms.

# Diagnosing Learning Algorithms with LRV

How to diagnose, e.g., convergence, in the previous setting?

**Tool: Central Limit Theorem**

Under suitable conditions, $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \sum_{k\in\mathbb{Z}} \gamma_k\right)$.

Long-run variance (LRV): $\sigma^2 = \sum_{k\in\mathbb{Z}} \gamma_k$

1. differs from sample variance $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ due to dependency;
2. needs to be updated sequentially to diagnose at different $n$.

**An Efficiency Dilemma with Existing Works**

1. Classical estimators: statistically efficient but $O(n)$-time update.
2. Recursive estimators: $O(1)$-time update but higher asymptotic mean squared error (AMSE).

# Our Contributions

As we investigate the efficiency dilemma, we develop and discuss:

1. (Theoretical) recursive LRV estimators with **super-optimal** AMSE as compared with their non-recursive counterparts;
2. (Theoretical) the first sufficient condition that characterizes $O(1)$-time or space updates;
3. (Computational) the **first mini-batch estimator** that can be much faster than existing algorithms (including recursive) in practice;
4. (Computational) automatic optimal parameters selection algorithm;
5. (Practical) applications in diagnosing Markov chain Monte Carlo (MCMC) and stochastic gradient descent (SGD).

**In the Poster . . .**

Points 1, 3 and 5 are discussed. The remaining parts need more elaboration and so deferred to the appendix here.

# Sneak Peek: Statistical Efficiency

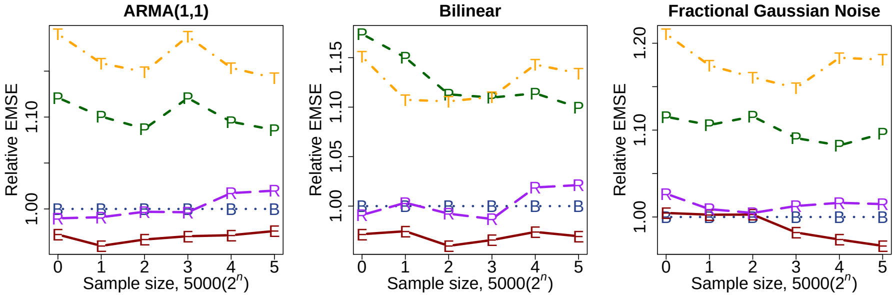


**Figure 1:** *Comparison of the relative empirical MSEs under Bartlett kernel ('B'), PSR ('P'), TSR ('T'), LASER(1,1) ('E') and LASER(1,2) ('R'). The experiments are conducted based on 1000 replications.*

# Sneak Peek: Computational Efficiency

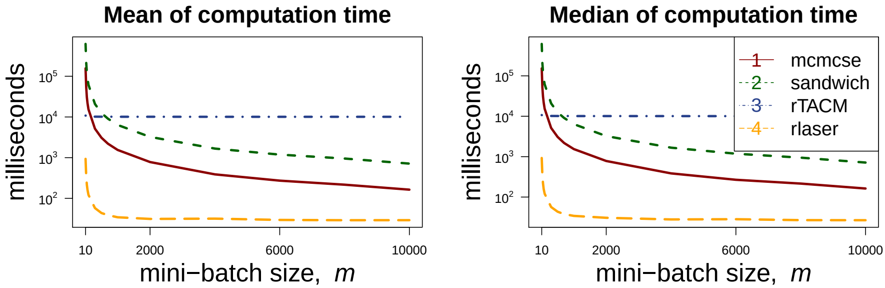


**Figure 2:** *Comparison of the computation time under existing implementations of Bartlett kernel (sandwich), overlapping batch means (mcmcse), PSR (rTACM) and mini-batch LASER (rlaser) in R. The experiment is conducted based on 50 replications and 100,000 samples.*

# Appendix

# Full Version of the LASER Principles

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} T\left(\frac{|i-j|}{t_n(i)}\right) S\left(\frac{|i-j|}{s_n(i)}\right) X_i X_j .$$

1. (**L**ocal Subsampling) An $O(1)$-time update algorithm should utilize local subsample.
2. (**A**synchronous Tapering) Under stationarity, $(X_i, X_j)$ and $(X_{i'}, X_{j'})$ should receive the same scaling if $|i-j| = |i'-j'|$.
3. (**S**eparated Parameters) The tapering and subsampling parameters should be separately chosen.
4. (**E**xterior Tapering) An $O(1)$-time update algorithm should exteriorize the tapering parameter.
5. (**R**amped Subsampling) An $O(1)$-space update algorithm should ramp up the subsample until it is too large.

# Time Complexity of $\hat{\sigma}_n^2$

## Sufficient Condition for $O(1)$-time Update

Let $q, C \in \mathbb{Z}^+$ and $c_0, \ldots, c_q \in \mathbb{R}$ be fixed. Suppose $\hat{\sigma}_n^2$ can be written as

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{j=1}^{n} T\left(\frac{|i-j|}{t_n}\right) S\left(\frac{|i-j|}{s_i}\right) X_j, \tag{1}$$

satisfying

1. the tapering function is of the form $T(u) = \sum_{r=0}^{q} c_r u^r$;
2. the subsampling function is of the form $S(u) = \mathbb{I}_{u<1}$;
3. the subsampling parameter $s_i$ is local and $|s_i - s_{i-1}| < C$.

Then $\hat{\sigma}_n^2$ can be updated in $O(1)$-time.

# Space Complexity of $\hat{\sigma}_n^2$

## Sufficient Condition for $O(1)$-space Update

Suppose $\hat{\sigma}_n^2$ can be written as (1), which satisfies

1. the estimator $\hat{\sigma}_n^2$ can be updated in $O(1)$-time;
2. the subsampling function is of the form $S(u) = \mathbb{I}_{u<1}$;
3. the ramped subsampling parameter $s_i'$ with $\phi \geq 2$ is used in place of $s_i$.

Then $\hat{\sigma}_n^2$ can be updated in $O(1)$-space.

# Automatic Optimal Parameters Selection

MSE-optimal parameters depend on $\kappa_q = |v_q|/\sigma^2$:

1. $\sigma^2$: readily available from last iteration
2. $v_q$: recursively estimated by extending LASER

$$\hat{v}_{n,\text{LASER}(1,\phi,1,q)} = \frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{s_i'-1} \left(1 - \frac{k}{t_n}\right) k^q X_i X_j.$$

Advantages of this extension:

1. Fully utilize available data as compared with pilot estimation.
2. Preserve desirable properties such as $O(1)$-space or mini-batch update.

# Models used in Monte Carlo Experiments

The following time series models are used:

1. *ARMA(1,1)*: Let $X_i - \mu = a(X_{i-1} - \mu) + b\varepsilon_{i-1} + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \nu^2)$. Take $a = 0.5$, $b = 0.5$, $\nu = 1$ and $\mu = 0$.
2. *Bilinear*: Let $X_i - \mu = (a + b\varepsilon_i)(X_{i-1} - \mu) + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \nu^2)$. Take $a = 0.9$, $b = 0.1$, $\nu = 1$ and $\mu = 0$.
3. *Fractional Gaussian Noise Process*: Let $X_i = Y_i$ be a zero-mean Gaussian processes with polynomial decaying ACVF, i.e., $\mathbb{E}(Y_0 Y_k) = a(k + b)^{-c}$. Take $a = 70$, $b = 7$ and $c = 3$.