Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Memoryless Cumulative Sign Detector for Stealthy CPS Sensor Attacks ⋆

Paul J Bonczek and Nicola Bezzo

Charles L. Brown Department of Electrical and Computer Engineering, and Link Lab, University of Virginia, Charlottesville, VA 22904, USA. (e-mail: {pjb4xn, nb6be}@virginia.edu).

Abstract: Stealthy false data injection attacks on cyber-physical systems introduce erroneous measurements onto sensors with the intent to degrade system performance. An intelligent attacker can design stealthy attacks with knowledge of the system model and noise characteristics to evade detection from state-of-the-art fault detectors by remaining within detection thresholds. However, during these hidden attacks, an attacker with the intention of hijacking a system will leave traces of non-random behavior that contradict with the expectation of the system model. Given these premises, in this paper we propose a run-time monitor called Cumulative Sign (CUSIGN) detector, for identifying stealthy falsified measurements by identifying if measurements are no longer behaving in a random manner. Specifically, our proposed CUSIGN monitor considers the changes in sign of the measurement residuals and their expected occurrence in order to detect if a sensor could be compromised. Moreover, our detector is designed to be a memoryless procedure, eliminating the need to store large sequences of data for attack detection. We characterize the detection capabilities of the proposed CUSIGN technique following the well-known χ^2 failure detection scheme. Additionally, we show the advantage of augmenting CUSIGN to the model-based Cumulative Sum (CUSUM) detector, which provides magnitude bounds on attacks, for enhanced detection of sensor spoofing attacks. Our approach is validated with simulations on an unmanned ground vehicle (UGV) during a navigation case study.

Keywords: Attack detection, Fault detection, Cyber-physical systems, Sensor spoofing

1. INTRODUCTION

Today's cyber-physical systems (CPSs) are fitted with multiple on-board sensors and computers that make them capable of many civilian and military applications with minimal/no human supervision. Autonomous navigation, transportation, surveillance, and task oriented jobs are becoming more common and ready for deployment in real world applications especially in the automotive, industrial, and military domains. These various enhancements in autonomy are possible thanks to the tight interaction between computation, sensing, communications, and actuation that characterize CPSs. With these increasing capabilities, comes the risk of more security vulnerabilities to cyber-attacks like sensor spoofing with the intent to induce undesired system behavior. An example of this problem was demonstrated by authors in [Bhatti and Humphreys (2017)] in which GPS data were spoofed to slowly drive a yacht off the intended route.

Many systems, including vehicle technologies, typically have well studied dynamical models and their sensors have specific expected behaviors according to their characterized noise profiles. Malicious attackers aim to compromise a system by diverting system states to unsafe regions, while remaining hidden within system detection boundaries. Despite lying within magnitude boundaries to remain undetected, non-random patterns arise that violate the expected behavior from normal system behavior. For example an attacker with the intention of hijacking an autonomous system while remaining stealthy will manipulate sensor measurements pushing them toward one direction.

Considering the problem at hand, in this work, we leverage the known characteristics of the *residual* – defined as the difference between sensor measurement and state prediction – to build a memoryless run-time monitor to detect non-random behaviors in sensing. To this end, we consider the χ^2 detection scheme [Mo et al. (2010)] which creates a *test measure* to monitor a vector of Normally distributed residuals. To monitor for randomness, we leverage the signed value of the difference between the χ^2 distributed test measure and an arbitrary reference point within its known distribution. Systems operating under normal conditions have expected probabilities of whether the test measure should be greater or less than the chosen reference point. Our Cumulative Sign (CUSIGN) dynamic detector, inspired by Cumulative Sum (CUSUM) theory [Page (1954)], leverages the history of sign valued differences between the test measure and the reference point, resulting in an alarm rate which is monitored at run-time for attack detection purposes. Thus, as a sensor is compromised, its corresponding residual will leave a trail of non-random behavior and will not follow an expectation.

In summary, the main objective of this work is to find stealthy sensor attacks exhibiting non-random behavior within the noise profile of a system in the presence of sensor and process noise. The contribution of this paper is twofold: 1) we propose the CUSIGN detection framework to deal with hidden non-random sensor attacks, typically undetectable by conventional detectors, by monitoring the expected alarm rate associated with consecutive changes

⋆ This work is based on research sponsored by ONR under agreement number N000141712012, and NSF under grant #1816591.

of signs in the *test measure*; 2) we introduce a memoryless feature to the CUSIGN detection procedure by leveraging a modified version of Welford's online algorithm [Welford (1962)], which we call a *Memoryless Run-time Estimator* (MRE), that uses a pseudo-window to monitor the alarm rate at run-time, removing the need of storing the entire sequence of data over the duration of the operation. We show empirical results about the MRE with a chosen pseudo-window length to find bounds for detection. Our framework is also combined with the CUSUM technique to create a complete detector framework. Furthermore, we include simulations on a UGV model to validate the proposed detection scheme.

1.1 Related Work

The subject of CPS security has garnered considerable interest in analyzing detection methods for stealthy sensor attacks that intend to degrade system performance. This work builds on previous research considering deceptive cyber-attacks to systems by injecting false data to sensor measurements while trying to remain undetected [Mo et al. (2010)]. Previous works have analyzed the effects of malicious sensor attacks on the Kalman filter [Bai and Gupta (2014)]. Similarly, authors in [Mo et al. (2010); Kwon et al. (2013)] discuss how undetected attacks can compromise closed-loop systems, causing state and system dynamic degradation.

Several attack detection techniques exist in the literature that analyze the residual, one of which is the Sequential Probability Ratio Testing (SPRT) [Kwon et al. (2016)] that tests the sequence of incoming residuals one at a time by taking the log-likelihood function (LLF). Compound Scalar Testing (CST) in [Kwon et al. (2013)] is a computationally friendly technique that reduces the residual vector with known residual variances into a scalar test measure of χ^2 distribution. An improvement of CST in [Miao et al. (2014)] is made by including a coding matrix to sensor outputs that is unknown to attackers, then an iterative optimization algorithm is used to solve for a transform matrix to detect stealthy attacks.

Different from these previous works that leverage residual-based techniques, we build a framework to monitor sensor measurements to find previously undetectable attacks by searching for non-random behavior. The CUSIGN detector proposed in this work to find non-random behaviors is inspired by the theory of CUmulative SUM (CUSUM), developed in [Page (1954)] that is commonly used as a monitor for change detection, such as a change in mean. Authors in [Murguia and Ruths (2019)] formalized a model-based detector of the CUSUM algorithm by leveraging known characteristics of the system dynamical and noise models.

Several statistical techniques are available in literature to test for randomness by leveraging a sequence of data and test a hypothesis. Among randomness tests, the Wald-Wolfowitz runs test [Wald and Wolfowitz (1940)] observes consecutive values that belong to one of two different groups (known as a *run*) over a given sequence. Similarly, the Serial Independence runs test [Cammarota (2011)] observes the number of runs of the difference between current and previous values over a sequence of data. The Monobit (frequency) test [Zhu et al. (2016)] observes a sequence of 1's and 0's to determine if they are equally probable. This work leverage these principles to detect non-random patterns in sensor measurements.

The remainder of this work is organized as follows: In Section 2 we begin by introducing the system modeling and problem formulation, followed by the characterization of our CUSIGN detector with an empirically derived memoryless detector operation to provide detection bounds in Section 3. In Section 4 we briefly discuss the CUSUM attack detector to compare with CUSIGN. Finally, in Section 5 we demonstrate through simulations the performance of our framework augmented with an existing CUSUM detector before drawing conclusions in Section 6.

2. PRELIMINARIES & PROBLEM FORMULATION

2.1 Model

In this work we consider autonomous cyber-physical systems whose dynamics are described by a discrete-time linear system in the following form:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\nu}_k \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k, \end{aligned} \tag{1}$$

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the state matrix, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ the input matrix, and $\boldsymbol{C} \in \mathbb{R}^{s\times n}$ the output matrix with the state vector $\boldsymbol{x}_k \in \mathbb{R}^n$, system input $\boldsymbol{u}_k \in \mathbb{R}^m$, output vector $\boldsymbol{y}_k \in \mathbb{R}^s$ providing measurements from s sensors from the set $\mathcal{S} = \{1, 2, \ldots, s\}$, and sampling time-instants $k \in \mathbb{N}$. Process and measurement noises are multivariate zero-mean Gaussian uncertainties $\boldsymbol{\nu} = \mathcal{N}(0, \boldsymbol{Q}) \in \mathbb{R}^n$ and $\boldsymbol{\eta} = \mathcal{N}(0, \boldsymbol{R}) \in \mathbb{R}^s$ with covariance matrices $\boldsymbol{Q} \in \mathbb{R}^{n\times n}, \boldsymbol{Q} \geq 0$ and $\boldsymbol{R} \in \mathbb{R}^{s\times s}, \boldsymbol{R} \geq 0$ respectively, and are assumed static.

During operations, a Kalman Filter (KF) is implemented to provide a state estimate $\hat{\boldsymbol{x}}_k \in \mathbb{R}^n$ in the following form,

$$\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{L}(\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k), \tag{2}$$

where the Kalman gain matrix $\boldsymbol{L} \in \mathbb{R}^{n\times s}$ is

$$\boldsymbol{L} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T + \boldsymbol{R})^{-1}. \tag{3}$$

For ease, we assume that the KF is at steady state before sensor attacks occur, such that $\lim_{k\to\infty} \boldsymbol{P}_k = \boldsymbol{P}$. The estimation error of the steady state KF is defined as $\boldsymbol{e}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$ while its *residual* $\boldsymbol{r}_k$ is given by

$$\boldsymbol{r}_k = \boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k = \boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k, \tag{4}$$

and the covariance matrix of the residual (4) is defined as

$$\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{r}_{k+1}\boldsymbol{r}_{k+1}^T] = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T + \boldsymbol{R} \in \mathbb{R}^{s\times s}. \tag{5}$$

2.2 Residual for Detection

A widely used sensor measurement failure detector in CPSs is the χ^2 detector [Mo et al. (2010)], computed by the following quadratic test measure

$$z_k = \boldsymbol{r}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_k = \boldsymbol{r}_k^T(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T + \boldsymbol{R})^{-1}\boldsymbol{r}_k \in \mathbb{R}^{\geq 0}. \tag{6}$$

In the absence of sensor attacks, the residual is a Normally distributed random vector $\boldsymbol{r}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{r}_k \in \mathbb{R}^s$, the test measure z_k belongs to a χ^2 distribution with s degrees of freedom. Under the assumption that the system is not under attack (i.e. the residual satisfies (5)), the scalar test measure in (6) follows

$$\mathrm{E}[z_k] = s, \quad \mathrm{Var}[z_k] = 2s. \tag{7}$$

The general case of the χ^2 detector compares the scalar test measure z_k to a threshold T by:

$$\begin{cases} z_k \leq T \longrightarrow \textit{no alarm}, \\ z_k > T \longrightarrow \textit{alarm}. \end{cases} \tag{8}$$

where the design of the threshold T is independent of system noises and based on the number of sensors, s.

When one or more sensors are attacked, properties of the residual $\boldsymbol{r}_k$ and the test measure z_k may no longer hold. When considering sensor attacks, the output of the system can be written as:

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k, \tag{9}$$

where $\boldsymbol{\xi}_k \in \mathbb{R}^s$ represents the attack vector subject to false data injection attacks onto sensors. A limitation to the method from (8) is that stealthy attacks purposely hidden within detection boundaries may be undetectable. However, an attacker with the intent of hijacking a CPS, may leave traces of non-random behavior on the test measure z_k. To detect such non-random behavior, we propose a framework consisting of adding a memoryless run-time dynamic detector on the test measure z_k searching for non-random behavior, while eliminating the need to store large amounts of data for detection purposes.

2.3 Problem Statement

An attacker trying to hijack a system, will consequently leave behind non-random behavior to sensor measurements. In this work we focus specifically on sign changes and their expected occurrences. With these considerations in mind, a system that is not compromised will have measurement residuals with signs that are normally distributed and with proper rate of sign changes.

***Definition** 1.* A system that is not compromised will behave in a random manner if the signed value of the difference between the test measure and a reference point maintain an expected sign occurrence.

Since we are considering sensor spoofing, unknown attack signals containing malicious data can disrupt randomness, resulting in measurements that display non-random signed behavior. Formally, the problem that we are interested in solving is:

***Problem** 1.* **Randomness of Measurements:** Given the test measure (6) computed from the residual $\boldsymbol{r}_k$ and the residual covariance $\boldsymbol{\Sigma}$ as defined in (4) and (5), find a policy to determine at run-time whether a sensor measurement is non-random, i.e., if the condition in Definition 1 does not hold.

Furthermore, in this work we impose that the computation from all aspects of the detector must have a memoryless property.

***Definition** 2.* A detector satisfies the memoryless property when the detection procedure does not rely on storing and using a sequence of data over any window horizon.

***Problem** 2.* **Memoryless Property:** With the given test measure (6) to analyze, find a policy for a memoryless detection procedure without the need to store a collection of data over any length of time to determine if the system is compromised, satisfying the condition in Definition 2.

2.4 System Architecture

The overall cyber-physical system architecture including the CUSIGN detector is summarized in Fig. 1. CUSIGN,

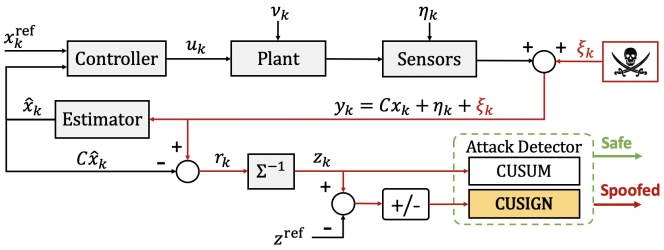


Fig. 1. The architecture of a CPS while experiencing sensor attacks augmented with our CUSIGN detector.

which can be augmented to any boundary detector providing magnitude bounds, is placed in the system feedback to monitor the relationship between measurement and state prediction. We focus on stealthy sensor attacks where an attacker may inject an attack signal at any point between the sensors and the state estimator, in an attempt to affect system behavior.

3. CUMULATIVE SIGN DETECTOR

We develop a Cumulative Sign (CUSIGN) detector that analyzes the sign of the given test measure z_k relative to a reference point and determines whether there is non-random behavior occurring. The model-based CUSIGN detector monitors the test measure from (6) and outputs an alarm when the CUSIGN test variable reaches a user defined threshold. For a given user defined threshold, an expected alarm rate can be found that is independent from the model of the system (1).

In normal conditions, i.e., without attacks or sensor malfunctions, the test measure z_k has a specific probability of being higher or lower than a given user defined reference point $z^{\text{ref}} \in \mathbb{R}^{>0}$ within its known distribution. We formalize these probabilities of z_k being higher or lower than the reference point by

$$\begin{aligned} \Pr\big(z_k < z^{\text{ref}}\big) &= \gamma\Big(\frac{s}{2}, \frac{z^{\text{ref}}}{2}\Big), \\ \Pr\big(z_k > z^{\text{ref}}\big) &= 1 - \gamma\Big(\frac{s}{2}, \frac{z^{\text{ref}}}{2}\Big), \end{aligned} \tag{10}$$

where $\gamma(\cdot,\cdot)$ is the *regularized lower incomplete gamma function* [Ross (2006)]. The sign of z_k with respect to the reference z^{ref} is computed by the following

$$\operatorname{sgn}(z_k - z^{\text{ref}}) := \begin{cases} -1, & \textbf{if } z_k - z^{\text{ref}} < 0, \\ 0, & \textbf{if } z_k - z^{\text{ref}} = 0, \\ 1, & \textbf{if } z_k - z^{\text{ref}} > 0, \end{cases} \tag{11}$$

where the probability of each scenario occurring is

$$\begin{aligned} \Pr\big(\operatorname{sgn}(z_k - z^{\text{ref}}) = -1\big) &= p_-, \\ \Pr\big(\operatorname{sgn}(z_k - z^{\text{ref}}) = 0\big) &= 0, \\ \Pr\big(\operatorname{sgn}(z_k - z^{\text{ref}}) = 1\big) &= p_+. \end{aligned} \tag{12}$$

An example of (12) is shown in Fig. 2, where the probabilities p_+ and p_- determine whether z_k will be higher or lower than z^{ref} given $z_k \sim \chi^2$.

The procedure of CUSIGN is an accumulation of signed values, denoted by the CUSIGN test variables S_k^+ and S_k^-. Each variable is a monitor checking for a change in the probability of the signed value $\operatorname{sgn}(z_k - z^{\text{ref}})$, one for *positive* and the other for *negative* changes. The following

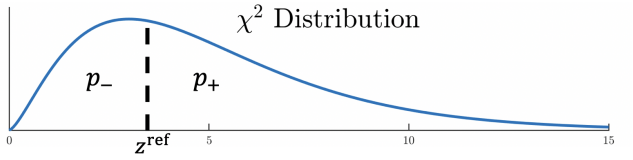


Fig. 2. Probabilities p_+ and p_- determined by z^{ref}.

procedure summarizes the CUSIGN detection in both the positive and negative cases:

CUSIGN Detector Procedure

$$\begin{aligned}
&\textbf{Initialize } S_0^+ = 0,\\
&S_k^+ = \max\big(0, S_{k-1}^+ + \operatorname{sgn}(z_k - z^{\text{ref}})\big),\\
&S_k^+ = 0 \text{ and Alarm } \zeta_k^+ = 1, \qquad \textbf{if } S_{k-1}^+ = \tau,\\
&\textbf{Initialize } S_0^- = 0,\\
&S_k^- = \min\big(0, S_{k-1}^- + \operatorname{sgn}(z_k - z^{\text{ref}})\big),\\
&S_k^- = 0 \text{ and Alarm } \zeta_k^- = 1, \qquad \textbf{if } S_{k-1}^- = -\tau.
\end{aligned} \tag{13}$$

The design of the test variable sequences S_k^+ and S_k^- are to accumulate the signed value $\operatorname{sgn}(z_k - z^{\text{ref}}) \in \{-1, 0, 1\}$ and triggering an alarm $\zeta_k^\pm = \{\zeta_k^+, \zeta_k^-\} \in \{0, 1\}$ when the test variables reach the threshold values $\tau \in \mathbb{N}^+$. When either of the test variables are equal to their corresponding thresholds, the given test variable is reset to 0. An example of the CUSIGN test variable is shown in Fig. 3 where three consecutive iterations $z_k > z^{\text{ref}}$ are satisfied at $k = 1, 2, 3$ (transitioning S_k^+ in the direction of p_+). At $k = 3$, the CUSIGN test variable S_k^+ reaches the threshold value $\tau = 3$ causing a reset such that $S_k^+ \to 0$.

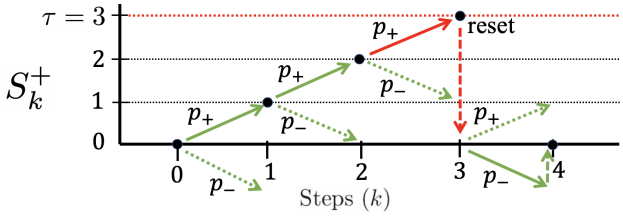


Fig. 3. Transitions of the CUSIGN test variable S_k^+ with threshold $\tau = 3$.

Choosing a specific threshold τ results in expected alarm rates $\mathrm{E}[\alpha^+]$ and $\mathrm{E}[\alpha^-]$ for both the positive and negative cases of the CUSIGN procedure (13). In the case that $z^{\text{ref}} = \mathrm{E}[\text{median}(z_k)]$ such that $p_+ = p_-$, the resulting expected alarm rates are equal $\mathrm{E}[\alpha^+] = \mathrm{E}[\alpha^-]$.

Similar to the implementation in [Murguia and Ruths (2019)], the transition of the CUSIGN test sequences $S_k^\pm$ can be constructed as a Markov chain with a transition matrix modeled from the probabilities of $\operatorname{sgn}(z_k - z^{\text{ref}})$. With a user defined threshold τ to trigger an alarm and causing a reset condition of the CUSIGN test variable to 0, we show the transitions of $S_k^\pm$ with a Markov chain diagram, as follows in Fig. 4.

Given a chosen threshold value $\tau \in \mathbb{N}^+$ as a value that triggers an alarm when $|S_k^\pm| = \tau$, we describe the Markov chain in Fig. 4 in the form of a Markov transition matrix $\mathcal{T}^\pm \in \mathbb{R}^{(\tau+1)\times(\tau+1)}$. The CUSIGN Markov Chain, occurring in a discrete manner, contains $\tau + 1$ states denoted as $\mathcal{M} = \{M_0, M_1, \ldots, M_\tau\}$ where M_τ is an absorbing state that is equal to the threshold, causing the CUSIGN test sequence $S_k^\pm$ to reset to M_0 (i.e., $S_k^\pm = 0$).

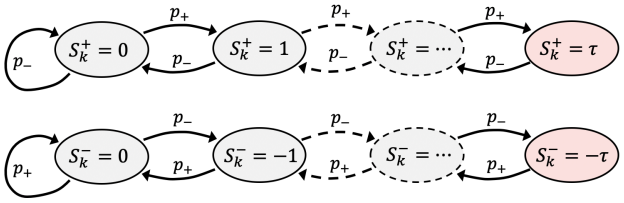


Fig. 4. Markov chain for both positive and negative cases of the CUSIGN test sequence with threshold τ.

The CUSIGN Markov transition matrix $\mathcal{T}^\pm$ for both positive $\mathcal{T}^+$ and negative $\mathcal{T}^-$ cases with a probability distribution of $\operatorname{sgn}(z_k - z^{\text{ref}})$ are written by

$$\mathcal{T}^\pm = \begin{bmatrix} p_\mp & p_\pm & 0 & 0 & \dots & 0 \\ p_\mp & 0 & p_\pm & 0 & \dots & 0 \\ 0 & p_\mp & 0 & p_\pm & & 0 \\ \vdots & & \ddots & & \ddots & \vdots \\ 0 & \dots & 0 & p_\mp & 0 & p_\pm \\ 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{14}$$

The transition matrix $\mathcal{T}^\pm$ structure remains the same on any system, where the matrix size depends only on the value of the threshold τ. Transition probabilities for transient states in $\mathcal{T}^\pm$ adhere to the following

$$\mathcal{T}^\pm \begin{cases} \Pr(M_j \to M_{j+1}) = p_\pm, \text{ for } j = \{0, \ldots, \tau-1\}, \\ \Pr(M_j \to M_{j-1}) = p_\mp, \text{ for } j = \{1, \ldots, \tau-1\}, \\ \Pr(M_0 \to M_0) = p_\mp, \end{cases} \tag{15}$$

and the final row represents an absorbing state containing elements equal to 0 besides the last element equaling 1.

We define $\mathcal{R}^\pm \in \mathbb{R}^{\tau\times\tau}$ as a matrix obtained from $\mathcal{T}^\pm$ with its last row and column removed (i.e., the absorbing state at threshold τ is removed), representing the transition probabilities to and from the transient states, also known as the fundamental matrix. Elements of $\mathcal{R}^\pm$ are all non-negative and row sums are equal to or less than one, while the eigenvalues satisfy $\rho[\mathcal{R}^\pm] < 1$ such that $(\mathcal{R}^\pm)^i \to 0$ as $i \to \infty$ and $\sum_{i=0}^{\infty}(\mathcal{R}^\pm)^i = (\boldsymbol{I}_\tau - \mathcal{R}^\pm)^{-1}$, where $\rho[\cdot]$ is the spectral radius and $\boldsymbol{I}_\tau$ is the identity matrix of size τ.

Lemma 1. Given a system with a CUSIGN detector (13) with a chosen threshold $\tau \in \mathbb{N}^+$ and reference point $z^{\text{ref}} \in \mathbb{R}^{>0}$ that is not affected by sensor attacks such that the residual sequence satisfies $\boldsymbol{r}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \in \mathbb{R}^s$ and $z_k = \boldsymbol{r}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_k \sim \chi^2$ with s degrees of freedom, then the inverse of the first element of the following vector

$$\boldsymbol{\mu}^\pm = (\boldsymbol{I}_\tau - \mathcal{R}^\pm)^{-1} \mathbf{1}_{\tau\times 1} = (\mu_1^\pm, \ldots, \mu_\tau^\pm)^T, \tag{16}$$

is the expected alarm rate, i.e., $\mathrm{E}[\alpha^\pm] = (\mu_1^\pm)^{-1}$.

Proof. Given the Markov chain containing $\tau+1$ states denoted by $\mathcal{M} = \{M_0, M_1, \ldots, M_\tau\}$, a fundamental matrix $\mathcal{R}^\pm$ is taken from a designed Markov transition matrix (14) to satisfy the transition probabilities (15). Leveraging the theory of average run length (ARL) in CUSUM [Brook and Evans (1972)], the ARL is defined as the average length of time for the test sequence to reach the threshold τ to trigger an alarm, determined by the fundamental matrix $\mathcal{R}^\pm$ containing the transient states within $\mathcal{T}^\pm$. By definition, the inverse of the ARL to observe an alarm results in the average frequency of obtaining an alarm, known as the alarm rate. The ARL can be found by computing (16), then by inverting the first element of $\boldsymbol{\mu}^\pm$, i.e., $(\mu_1^\pm)^{-1}$, we obtain the expected alarm rate $\mathrm{E}[\alpha^\pm]$.

3.1 Memoryless Run-time Estimation of Alarm Rates

In the design of CUSIGN, we trigger an alarm when a test variable reaches a chosen threshold τ. Given a system not experiencing sensor attacks, we have an expectation of the alarm rates. Typically, to find an alarm rate, the number of triggered alarms are tallied over a given period of time. In this work, we want to create a "memoryless" procedure to find an alarm rate.

The conventional method of finding an average $\bar{x}$ of a stochastic variable is $\bar{x}_n = \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]$ where n is the size the data set. This procedure requires storage of the complete data set, where computation becomes less efficient as n grows. A memoryless online algorithm known as Welford's online algorithm for computing a mean incrementally was developed in [Welford (1962)] by transforming the conventional method into an online update by the following form

$$\begin{aligned}\bar{x}_n &= \frac{1}{n}\left[x_n + \sum_{i=1}^{n-1} x_i\right] = \frac{1}{n}\left[x_n + (n-1)\bar{x}_{n-1}\right] \\ &= \frac{1}{n}\left[x_n + n\bar{x}_{n-1} - \bar{x}_{n-1}\right] = \bar{x}_{n-1} + \frac{\left[x_n - \bar{x}_{n-1}\right]}{n}.\end{aligned} \tag{17}$$

It can be seen in (17) that n grows indefinitely, equal to the number of data points. We set a maximum value for n such that $\max(n) = \ell \in \mathbb{N}^+$ to create a "pseudo-window" for a rolling sequential estimation of an expected mean. We name this modified version of Welford's online algorithm utilizing a pseudo-window ℓ as a Memoryless Run-time Estimator (MRE). The behavior of MRE when computing the mean similarly imitates the conventional method of calculating the mean consisting of ℓ data points, but without the need to store the entire sequence.

For the case of attack detection using alarm rates for CUSIGN, we leverage MRE in (17) to find an online estimation of an expected alarm rate $\mathrm{E}[\alpha]$ (we omit $\pm$ for $\alpha^{\pm}$ in this section as the MRE applies to both the positive and negative cases). Leveraging the pseudo-window of length ℓ and replacing the counter n from (17) with k for sampling time instances, we attain the equation

$$\hat{\alpha}_k = \hat{\alpha}_{k-1} + \frac{\left[\zeta_k - \hat{\alpha}_{k-1}\right]}{\ell}, \tag{18}$$

where ζ_k is the triggered alarm for CUSIGN, $\hat{\alpha}_k$ is an estimate of the alarm rate at time instance k, and $\hat{\alpha}_0 = 0$ initially at $k = 0$.

***Proposition** 1.* Assuming the system is not experiencing sensor attacks and the test measure follows $z_k \sim \chi^2$ for time instances $k \geq 0$, we empirically find that the alarm rate is a Normal distribution as follows

$$\hat{\alpha} \sim \mathcal{N}\left(\mathrm{E}[\alpha], \frac{\theta \mathrm{E}[\alpha](1-\mathrm{E}[\alpha])}{\ell}\right), \tag{19}$$

where ℓ is the user defined pseudo-window length, $\theta \in \mathbb{R}^{>0}$ is an empirically found scaling value, and $\mathrm{E}[\alpha]$ is the expected alarm rate, i.e., the probability that the test variable S_k reach the threshold, triggering an alarm $\zeta_k = 1$.

Given the distribution of $\hat{\alpha}$ in Proposition 1, the expectation of the estimated alarm rate follows

$$\mathrm{E}[\hat{\alpha}] = \mathrm{E}[\alpha], \quad \mathrm{Var}[\hat{\alpha}] = \frac{\theta \mathrm{E}[\alpha](1-\mathrm{E}[\alpha])}{\ell}. \tag{20}$$

Values of θ are found to be dependent on the chosen threshold τ. Observed approximates of θ are presented in Table 1 for thresholds $\tau = 1, 2, 3, 4$ and $\ell \geq 10$.

Table 1. Empirical values for the scaling value θ given thresholds $\tau = 1, 2, 3, 4$.

Thresholds	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$
θ	$\frac{\ell}{2\ell-1}$	$\frac{.74\ell}{2\ell-1}$	$\frac{.7\ell}{2\ell-1}$	$\frac{.69\ell}{2\ell-1}$

***Remark** 1.* For the CUSIGN detector, we empirically find that $\hat{\alpha}$ follows (19) when $p_+ \approx p_-$ (i.e., z^{ref} is chosen to be at $\mathrm{E}[\mathrm{median}(z_k)]$ such that $p_- = p_+ = 0.5$). For a reference point z^{ref} not placed near the expected distribution median, i.e., $p_+ \not\approx p_-$, we found that the distribution of $\hat{\alpha}$ loses properties of the Normal distribution in (19). Empirical results for observed $\hat{\alpha}$ and $\mathrm{Var}[\hat{\alpha}]$ considering the case when $p_+ \not\approx p_-$ can be found in Appendix A.

By leveraging the distribution of the estimated alarm rate in (19), bounds of the alarm rate can be made.

***Lemma** 2.* Assuming an uncompromised system with a CUSIGN detector (13) with a reference point z^{ref} and threshold $\tau \in \mathbb{N}^+$, detection of sensor attacks occurs when $\tau_-^{\alpha} \leq \hat{\alpha} \leq \tau_+^{\alpha}$ where

$$\tau_{\pm}^{\alpha} = \mathrm{E}[\alpha] \pm Z\sqrt{\frac{\theta \mathrm{E}[\alpha](1-\mathrm{E}[\alpha])}{\ell}}. \tag{21}$$

Proof. Given a CUSIGN detector with threshold $\tau \in \mathbb{N}^+$ and reference point $z^{\mathrm{ref}} \in \mathbb{R}^{>0}$ that determine transition probabilities p_- and p_+, an expected alarm rate $\mathrm{E}[\alpha]$ can be computed by inverting the first element in (16). With $\mathrm{E}[\alpha]$ and leveraging the Memoryless Run-time Estimator with a pseudo-window of length ℓ, the distribution of the estimated alarm rate follows $\hat{\alpha} \sim \mathcal{N}(\cdot, \cdot)$ with properties from (20). Detection bounds $\tau_{\pm}^{\alpha}$ of a specific confidence level determined by Z of a Normally distributed random variable with properties from (20) follow $\mathrm{E}[\alpha] - Z\sqrt{\frac{\theta \mathrm{E}[\alpha](1-\mathrm{E}[\alpha])}{\ell}} \leq \hat{\alpha} \leq \mathrm{E}[\alpha] + Z\sqrt{\frac{\theta \mathrm{E}[\alpha](1-\mathrm{E}[\alpha])}{\ell}}$ satisfying (21), concluding the proof.

Detection of sensor attacks occur when an estimated alarm rate $\hat{\alpha}$ goes beyond a threshold from $\tau_{\pm}^{\alpha} = \{\tau_-^{\alpha}, \tau_+^{\alpha}\}$. Lower bounds resulting in $\tau_-^{\alpha} < 0$ are omitted as $\hat{\alpha} \in [0, \frac{1}{\tau}]$.

4. CUSUM DETECTOR REVIEW

The CUSIGN detector alone may not be sufficient as an attacker can change the magnitude of a measurement, but still maintain random signed behavior of the test measure z_k. The non-parametric quality of CUSIGN results in the inability to monitor the magnitude of the test measure. A well-known dynamic detector, the *CUmulative SUM* (CUSUM) detector, leverages the magnitude of the test measure sequence z_k to look for changes in the mean from an expectation. Formalized into a model-based attack detector by [Murguia and Ruths (2019)] that outputs an alarm, the CUSUM attack detection procedure follows

CUSUM Detector Procedure

$$\begin{aligned}&\textbf{Initialize } C_0 = 0, \\ &C_k = \max(0, C_{k-1} + z_k - b), \quad \textbf{if } C_{k-1} \leq \tau^C, \\ &C_k = 0 \text{ and Alarm } \zeta_k^C = 1, \quad \textbf{if } C_{k-1} > \tau^C.\end{aligned} \tag{22}$$

The working principle of this detector is to accumulate the test measure (6) in C_k, triggering an alarm $\zeta_k^C = 1$

when the test variable surpasses the threshold τ^C. The test variable C_k resets to zero either when the threshold τ^C is surpassed or when C_k goes negative. A bias b is selected based on properties of (5) such that C_k does not grow unbounded. A detailed explanation of how to construct a transition matrix for the probability distribution $z_k - b$ for the model-based CUSUM can be found in [Murguia and Ruths (2019)]. The authors provide a method for tuning the threshold τ^C given a bias b for a desired alarm rate $\mathrm{E}[\alpha^C]$ with an assumption that the system is free of sensor attacks, where the residual follows $\boldsymbol{r}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, hence a shifted χ^2 distribution $z_k - b = \boldsymbol{r}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_k - b$.

Considering CUSUM as a stand-alone detector, an adversarial wants to avoid attacks such that the test variable C_k exceeds threshold τ^C at a higher rate, thereby causing a reset $C_k = 0$ in (22) by satisfying the CUSUM procedure sequence $C_k = \max(0, C_{k-1} + z_k - b) \leq \tau^C$ to trigger alarms more often, resulting in a higher alarm rate α^C. Subsequently, an attacker can design an attack such that it remains within bounds of CUSUM to not trigger alarms more than expected. To include this attack vector, we can rewrite the CUSUM procedure such that

$$C_k = \max(0, C_{k-1} + \left(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k)\|^2\right) - b). \quad (23)$$

Assuming that a malicious attacker can have access to the sensor measurements $\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k$ and has perfect knowledge of the state estimator, it will be able to find the estimator output $\boldsymbol{C}\hat{\boldsymbol{x}}_k$. With this information, the attacker can solve for $\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k = \boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k$ to achieve the ability of manipulating elements of $\boldsymbol{\xi}_k$ by

$$\boldsymbol{\xi}_k = -\boldsymbol{C}\boldsymbol{e}_k - \boldsymbol{\eta}_k + \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\xi}_k^{\tau^C} \quad (24)$$

such that $\max(0, C_{k-1} + (\boldsymbol{\xi}_k^{\tau^C})^T \boldsymbol{\xi}_k^{\tau^C} - b) \leq \tau^C$ can maintain the test variable C_k within the detection threshold τ^C.

5. RESULTS

The proposed CUSIGN detector was validated in simulation and augmented with CUSUM introduced in Section 4. The case study presented in this paper is an autonomous way-point navigation of a skid-steering differential-drive UGV with the following linearized model [Nutaro (2011)]:

$$\begin{aligned} \dot{v} &= \frac{1}{m}(F_l + F_r - B_r v), \\ \dot{\omega} &= \frac{1}{I_z}\Big(\frac{w}{2}(F_l - F_r) - B_l \omega\Big), \ \dot{\theta}_h = \omega, \end{aligned} \quad (25)$$

where v, θ_h, and ω denotes the velocity, heading angle, and angular velocity, forming the state vector $\boldsymbol{x} = [v, \theta_h, \omega]^T$. F_l and F_r describe the left and right input forces from the wheels, w is the vehicle width, while B_r and B_l are resistances due to the wheels rolling and turning. The continuous-time model (25) is discretized with a sampling rate $t_s = 0.01$ to satisfy the system model described in (1). The UGV is tasked to continuously navigate to four goal-points along a square trajectory with side lengths of 5m maintaining a velocity $v = 0.5$m/s for 200s.

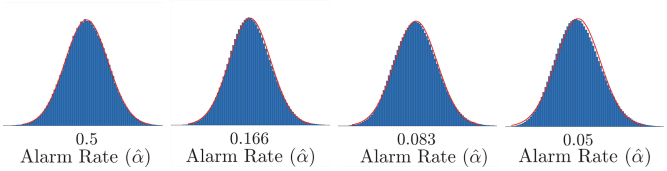
In the simulation, we perform two different attack sequences on the velocity sensor on-board the vehicle: 1) a persistent attack and 2) an alternating pattern attack. Both stealthy attack sequences are designed to be undetectable by CUSUM, but are detected by CUSIGN due to the creation of non-random patterns.

5.1 Simulations

We first consider the system under normal conditions where $\boldsymbol{\xi}_k = 0$. In Table 2 we show the alarm rate of the system over 5 million data samples and compare the results to the expected alarm rate $\mathrm{E}[\alpha^{\pm}]$ computed from (16) in the case where $p_+ = p_- = 0.5$ for thresholds $\tau = 1, 2, 3, 4$. Next, in Fig. 5 we show the distribution of the alarm rate estimate $\hat{\alpha}$ from the four cases in Table 2 overlaid with the expected distributed curve (in red) according to (20).

Table 2. $\mathrm{E}[\alpha^{\pm}]$ when $p_+ = p_- = 0.5$.

Thresholds	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$
$\mathrm{E}[\alpha^{\pm}]$	0.5	$0.1\bar{6}$	$0.08\bar{3}$	0.05
$\alpha^{\pm}$ (sim.)	.50006	.16692	.083291	.050012



(a) (b) (c) (d)

Fig. 5. Resulting distributions of $\hat{\alpha}$ when $p_+ = p_- = 0.5$ for (a) $\tau = 1$, (b) $\tau = 2$, (c) $\tau = 3$, (d) $\tau = 4$.

Now, considering the UGV (25) case study in the presence of hidden attacks on the velocity sensor on state $x_1 = v$, we show the detection capabilities of CUSIGN. The CUSIGN is designed with $z^{\mathrm{ref}} = \mathrm{E}[\mathrm{median}(z_k)] \approx s\left(1 - \frac{2}{9s}\right)^3$ where $s = 3$ such that the transition probabilities satisfy $p_{\pm} = 0.5$ and threshold $\tau = 2$. The expected alarm rate $\mathrm{E}[\alpha] = 0.1\bar{6}$ and the Memoryless Run-time Estimator (18) with pseudo-window length $\ell = 100$ has detection bounds (21) at $\tau_-^{\alpha} = 0.0987$ and $\tau_+^{\alpha} = 0.2347$ where $Z = 3$ for a 99.7% confidence. The design of CUSUM contains a bias $b = 1.1s = 3.3$ with a threshold $\tau^C = 2.3226$ to satisfy an expected alarm rate $\mathrm{E}[\alpha^C] = 0.15$ (see [Murguia and Ruths (2019)] for tuning details), where the alarm rate is computed by a conventional method of length ℓ by $\frac{1}{\ell}\sum_{k-\ell+1}^{k} \zeta_k^C$. Fig. 6 shows the results of a persistent attack (23), (24) beginning at $k = 10,000$ with a noiseless magnitude of $0.1\tau^C$. The alarm rate $\hat{\alpha}^C$ for CUSUM is unaffected while CUSIGN discovers the attack and alarm rates $\hat{\alpha}^{\pm}$ both go beyond the detection bounds $\tau_{\pm}^{\alpha}$ (red dashed lines). A second attack shown in Fig. 7 is attempted with an alternating noiseless pattern of $\{0.1\tau^C, -0.1\tau^C\}$ to show that CUSIGN can detect patterns. Again, alarm rates for CUSIGN find the non-random patterns and go beyond the detection bounds $\tau_{\pm}^{\alpha}$ while CUSUM is not able to detect the non-random behavior.

6. CONCLUSIONS & FUTURE WORK

In this paper we have characterized the CUSIGN procedure for detection of hidden sensor attacks that present non-random behavior. In particular, we have constructed a Markov chain of the CUSIGN test sequence to model a resulting expected alarm rate. We have formalized a memoryless run-time method for computing an alarm rate estimate using a modified version of Welford's online algorithm with a pseudo-window, which we call the Memoryless Run-time Estimator (MRE). We empirically found the resulting estimated alarm rate distribution and leveraged

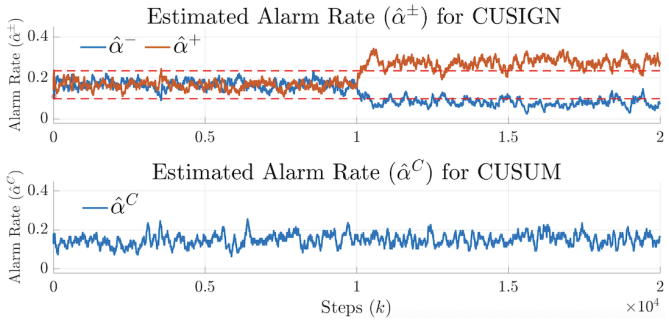


Fig. 6. Alarm rates $\hat{\alpha}^{\pm}$ and $\hat{\alpha}^C$ for both CUSIGN and CUSUM with a hidden persistent sensor attack.

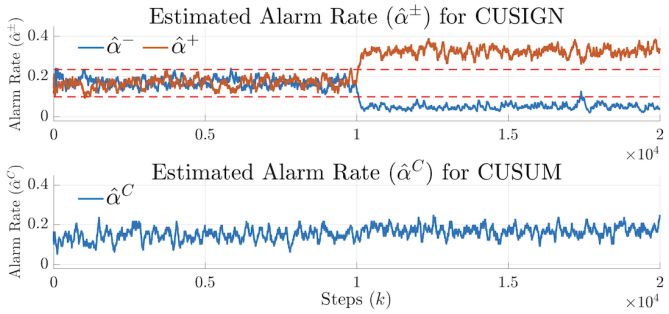


Fig. 7. Alarm rates $\hat{\alpha}^{\pm}$ and $\hat{\alpha}^C$ for both CUSIGN and CUSUM with a hidden alternating sensor attack.

it to provide detection bounds given a specific level of confidence. Then, we characterized attack sequences that remain undetected to the CUSUM dynamic attack detector, that leave trails of non-random behavior for CUSIGN to detect the attack.

In our future work we plan to extend the current work to leverage CUSIGN on CPSs with redundant sensors to detect and remove compromised sensors and create an attack resilient controller.

REFERENCES

Bai, C. and Gupta, V. (2014). On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds. In *2014 American Control Conference*.

Bhatti, J. and Humphreys, T.E. (2017). Hostile control of ships via false gps signals: Demonstration and detection. *Navigation*, 64(1), 51–66.

Brook, D. and Evans, D.A. (1972). An approach to the probability distribution of cusum run length. *Biometrika*, 59(3), 539–549.

Cammarota, C. (2011). The difference-sign runs length distribution in testing for serial independence. *Journal of Applied Statistics*, 38(5), 1033–1043.

Kwon, C., Liu, W., and Hwang, I. (2013). Security analysis for cyber-physical systems against stealthy deception attacks. In *2013 American Control Conference*.

Kwon, C., Yantek, S., and Hwang, I. (2016). Real-time safety assessment of unmanned aircraft systems against stealthy cyber attacks. *Journal of Aerospace Information Systems*, 13(1), 27–45.

Miao, F., Zhu, Q., Pajic, M., and Pappas, G.J. (2014). Coding sensor outputs for injection attacks detection. In *53rd IEEE Conference on Decision and Control*.

Mo, Y., Garone, E., Casavola, A., and Sinopoli, B. (2010). False data injection attacks against state estimation in wireless sensor networks. In *2010 IEEE 49th Conference on Decision and Control*, 5967–5972.

Murguia, C. and Ruths, J. (2019). On model-based detectors for linear time-invariant stochastic systems under sensor attacks. *IET Control Theory Applications*, 13(8), 1051–1061.

Nutaro, J.J. (2011). *Building software for simulation: theory and algorithms, with applications in C++*. John Wiley & Sons.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.

Ross, S.M. (2006). *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., Orlando, FL, USA.

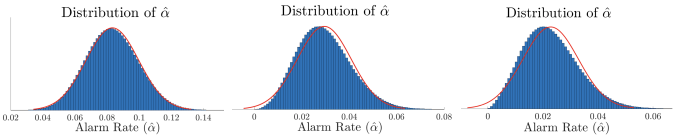
Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.*, 11(2), 147–162. doi:10.1214/aoms/1177731909.

Welford, B.P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3), 419–420.

Zhu, S., Ma, Y., Lin, J., Zhuang, J., and Jing, J. (2016). More powerful and reliable second-level statistical randomness tests for nist sp 800-22. In *ASIACRYPT*.

Appendix A. EMPIRICAL RESULTS

From Remark 1 in Section 3.1, we show in Fig. A.1 the gradual divergence from the normal approximation as p_+ and p_- are no longer similar as the distribution of the estimated alarm rate estimate $\hat{\alpha}$ becomes skewed. The empirical results provided throughout this section are results from 5 million samples, thus giving an accurate representation of the resulting distributions.



(a) (b) (c)

Fig. A.1. Resulting distributions of $\hat{\alpha}$ when (a) $p_+ = p_- = 0.5$, (b) $p_- = 0.37$, (c) $p_- = 0.3$.

Furthermore, Table A.1 provides the expected $\mathrm{E}[\hat{\alpha}]$ and simulated alarm rates, while Table A.2 provides the square root of the expected and simulated variance $\sqrt{\mathrm{Var}[\hat{\alpha}]}$ (i.e., standard deviation). It can be seen that as $p_{\pm} \approx 0.5$, the simulated mean of the alarm rate estimates remain approximately equal to the expectation (i.e., $\bar{\hat{\alpha}} \approx \mathrm{E}[\alpha]$), but the simulation results for standard deviation diverge from the expected variance as $p_+ \not\approx p_-$.

Table A.1. Results of $\mathrm{E}[\hat{\alpha}]$ for $\ell = 100$.

$p_{\pm}$	.4	.5	.6
$\mathrm{E}[\hat{\alpha}]$/sim (τ=1)	.400/4.01	.500/.500	.600/6.01
$\mathrm{E}[\hat{\alpha}]$/sim (τ=2)	.1143/.1142	.166$\bar{6}$/.1665	.2250/.2251
$\mathrm{E}[\hat{\alpha}]$/sim (τ=3)	.0484/.0483	.083$\bar{3}$/.0832	.1256/.1258
$\mathrm{E}[\hat{\alpha}]$/sim (τ=4)	.0244/.0239	.0500/.0500	.0835/.0833

Table A.2. Results of $\mathrm{std}[\hat{\alpha}]$ for $\ell = 100$.

$p_{\pm}$	.4	.5	.6
std[$\hat{\alpha}$]/sim (τ=1)	.0346/.0347	.0354/.0355	.0346/.0347
std[$\hat{\alpha}$]/sim (τ=2)	.0194/.0204	.0227/.0226	.0254/.0238
std[$\hat{\alpha}$]/sim (τ=3)	.0127/.0138	.0163/.0163	.0196/.0185
std[$\hat{\alpha}$]/sim (τ=4)	.0091/.0099	.0128/.0128	.0163/.0153