# Covariance-Robust Dynamic Watermarking

Matt Olfat, Stephen Sloan, Pedro Hespanhol, Matt Porter, Ram Vasudevan, and Anil Aswani

*Abstract*— **Attack detection and mitigation strategies for cyberphysical systems (CPS) are an active area of research, and researchers have developed a variety of attack-detection tools such as dynamic watermarking. However, such methods often make assumptions that are difficult to guarantee, such as exact knowledge of the distribution of measurement noise. Here, we develop a new dynamic watermarking method that we call covariance-robust dynamic watermarking, which is able to handle uncertainties in the covariance of measurement noise. Specifically, we consider two cases. In the first this covariance is fixed but unknown, and in the second this covariance is slowly-varying. For our tests, we only require knowledge of a set within which the covariance lies. Furthermore, we connect this problem to that of algorithmic fairness and the nascent field of fair hypothesis testing, and we show that our tests satisfy some notions of fairness. Finally, we exhibit the efficacy of our tests on empirical examples chosen to reflect values observed in a standard simulation model of autonomous vehicles.**

## I. INTRODUCTION

The development of 5G, the "fifth generation" of wireless technology, brings with it increased bandwidth, massive-scale device-to-device (D2D) connections, lower latency, and high reliability. The latency reductions with 5G open the door to further growth in cyberphysical systems (CPS), which involve the intercommunication and real-time management of large numbers of physical sensors and actuators, often in shifting environments [1]. System vulnerabilities to malicious agents abound in all of these technologies [2]–[4], and 5G in particular necessitates more robust cyber-security measures for the relevant control systems [1].

Much of the existing work on security for CPS assumes that the system is fixed and all required distributions are exactly known [5]–[7]. However, the system description is often time-varying or partially uncertain for many CPS [8], [9]. Given this real-world motivation of security for time-varying or partially unknown CPS, we focus in this paper on designing robust security schemes to test for adversarial attacks on LTI systems. Recent work has established dynamic watermarking as a key active method for detecting sensor attacks [5]–[7], [10]–[14], and here we build on this work by designing covariance-robust dynamic watermarking.

We design robust watermarking for two sub-cases: The first is where the covariance of measurement noise is fixed but unknown, and the second is where the covariance of measurement noise is unknown and slowly-varying. The first reflects a scenario of many nearly-identical systems with variation between copies of the system. The second sub-case reflects a scenario where a sensor has different accuracy in varying regimes, such as lidar on an autonomous vehicle in changing weather. Attack detection is critical in all of these such cases, and we need statistical tests that retain their power in the face of system changes or uncertainty.

### A. Fairness

Robust data-driven decision-making has gained attention in the literature on algorithmic fairness. Motivated by machine learning tasks with societal applications, the fairness literature has sought to design learning methods that refrain from considering certain variables. To that extent, this body of work defines rigorous, mathematical notions of fairness for supervised learning [15]–[22], which have recently been extended to unsupervised learning by [23], [24].

The work in [25] outlines a general framework: Consider $(X, Y, Z)$ with a joint distribution $\mathbb{P}$, where $X$ are exogenous inputs, $Y$ are endogenous "targets", and $Z$ is a "protected attribute". The goal is to choose a *decision rule* $\delta(x)$ that makes a decision $d$ using inputs $X$, in order to minimize some *risk function* $\mathcal{R}_{\mathbb{P}}(\delta, Y)$. In dynamic watermarking: $X$ are measurements, $Y$ is a binary variable that denotes if the system is under attack, and $Z$ is the true system characterization; our decision rule $\delta$ for if the system is under attack is made without $Y$ and $Z$, which are not observed. We then define a decision rule to be without *disparate impact* if

$$\delta^* \in \arg\min_{\delta} \big\{ R_{\mathbb{P}}(\delta, Y) \bigm| \delta(X) \perp\!\!\!\perp Z \big\}, \qquad (1)$$

where $\delta(X) \perp\!\!\!\perp Z$ means $\delta(X)$ is independent of $Z$. This increases fairness because it removes any impact of $Z$ on the decision by imposing independence as a constraint. However, some [16], [18] have argued that this above definition of fairness can be too restrictive in some cases and that *equalized odds* is a better definition of fairness. Its only difference is that in (1) we replace $\delta(X) \perp\!\!\!\perp Z$ with $(\delta(X) \perp\!\!\!\perp Z)|Y$. That is, equalized odds ask for independence of $\delta(X)$ and $Z$ when conditioned on $Y$. We can interpret equalized odds as requiring error rates to be similar across protected groups. Finally, a notion associated equalized odds is that of *equal opportunity*, which amounts to enforcing $(\delta(X) \perp\!\!\!\perp Z)|Y = y$, for some value $y$. This is relevant when one particular type of error is of more interest than another.

This material is based upon work partially supported by the National Science Foundation under Grant CMMI-1847666, by the UC Berkeley Center for Long-Term Cybersecurity, and by a grant from Ford Motor Company via the Ford-UM Alliance under award N022977.

Matt Olfat, Steve Sloan, Pedro Hespanhol, and Anil Aswani are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley 94720 {molfat,stephen_sloan,pedrohespanhol,aaswani}@berkeley.edu

Matt Porter and Ram Vasudevan are with the Department of Mechanical Engineering, University of Michigan, Ann Arbor 48109 {matthepo,ramv}@umich.edu

### B. Relevance of Fairness to Watermarking

Fairness is relevant to the design of robust tests for two reasons. First, it provides a well-established technical language with which to discuss our requirement of robustness. Past dynamic watermarking techniques require exact system knowledge, and as such the corresponding watermarking tests will have error rates that are biased over inevitable system perturbations or uncertainties. Fairness notions such as *equalized odds* and *equal opportunity* allow for more specific framing of the problem and thus give a framework to design more robust methods for dynamic watermarking.

Second, robust cyber-security methods will have improved social impacts, which is the most general way of interpreting "fairness". For example, smart homes can have many sensors. Changes in the distribution of sensor noise can correlate with factors such as climate, which correlates with geography and thus attributes like race, ethnicity, or class. A systemic bias in the ability to detect threats thus yields, and possibly perpetuates, systemic bias in outcomes among these groups. Robustness of cyber-security methods thus have the potential to improve societal fairness of the corresponding methods.

### C. Outline

In Sect. II, we outline key terminology and results in dynamic watermarking. In Sect. III-A, we present our covariance-robust dynamic watermarking scheme for the case of fixed, but unknown, measurement noise covariance. This is then extended in Sect. III-B to the case where measurement noise covariance is allowed to slowly vary. Sect. IV presents empirical results that demonstrate efficacy of our approach.

## II. Preliminaries

We describe the LTI system and attack models, and then review existing results about dynamic watermarking.

### A. LTI System Model

Consider a partially-observed MIMO LTI system

$$\begin{aligned} x_{n+1} &= Ax_n + Bu_n + w_n \\ y_n &= Cx_n + z_n + v_n \end{aligned} \tag{2}$$

for $x_n, w_n \in \mathbb{R}^p, u_n \in \mathbb{R}^q$ and $y_n, z_n, v_n \in \mathbb{R}^m$. Here $w_n$ is mean-zero i.i.d. multivariate Gaussian process noise with covariance matrix $\Sigma_W$, and this is independent of $z_n$ that is i.i.d. Gaussian measurement noise with mean-zero; but we assume that the covariance matrix for $z_n$ is a linear function $\Sigma_Z(\theta)$ of a set of parameters $\theta \in \mathcal{P} \subset \mathbb{R}^d$ taking values in polyhedron $\mathcal{P}$. For now, $\theta$ is assumed constant but unknown for any fixed system. The $v_n$ is an additive signal chosen by an attacker who seeks to corrupt sensor measurements.

Stabilizability of $(A, B)$ and detectability of $(A, C)$ imply the existence of a controller $K$ and observer $L$ such that $A + BK$ and $A + LC$ are Schur stable. The closed-loop system can be stabilized using the control input $u_n = K\hat{x}_n$, where $\hat{x}_n$ is the observer-estimated state. Define $\tilde{x}_n = \begin{bmatrix} x_n^\mathsf{T} & \hat{x}_n^\mathsf{T} \end{bmatrix}^\mathsf{T}$, $\underline{D} = \begin{bmatrix} I & 0 \end{bmatrix}^\mathsf{T}$, $\underline{L} = \begin{bmatrix} 0 & -L^\mathsf{T} \end{bmatrix}^\mathsf{T}$, and

$$\underline{A} = \begin{bmatrix} A & BK \\ -LC & A + BK + LC \end{bmatrix}. \tag{3}$$

We can write the closed-loop evolution of the state and estimated state when $v_n \equiv 0$ as $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{D}w_n + \underline{L}z_n$. Alternatively, we may define the observation error $\delta_n = \hat{x}_n - x_n$. Let $\breve{x}_n = \begin{bmatrix} x_n^\mathsf{T} & \delta_n^\mathsf{T} \end{bmatrix}^\mathsf{T}$, $\underline{\underline{D}} = \begin{bmatrix} I & -I \end{bmatrix}^\mathsf{T}$, $\underline{\underline{L}} = \underline{L}$, and

$$\underline{\underline{A}} = \begin{bmatrix} A + BK & BK \\ 0 & A + LC \end{bmatrix}. \tag{4}$$

The closed-loop system for this change of variables is $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_n + \underline{\underline{D}}w_n + \underline{\underline{L}}z_n$. Note that $\underline{\underline{A}}$ is Schur stable since both $A + BK$ and $A + LC$ are Schur stable.

### B. Attack Model

Following [26], we consider attacks where $v_n = \alpha(Cx_n + z_n) + C\eta_n + \zeta_n$ for a fixed $\alpha \in \mathbb{R}$ and i.i.d. Gaussian $\zeta_n$ with mean-zero and covariance matrix $\Sigma_S$. Here, the $\eta_n$ are chosen to follow the process $\eta_{n+1} = (A + BK)\eta_n + \omega_n$, where $\omega_n$ are similarly i.i.d. Gaussian with mean-zero and covariance matrix $\Sigma_O$. The implication is that the attacker minimizes or mutes the true output $Cx_n + z_n$, and instead replaces it with a simulated output that follows the system dynamics and is thus not easily distinguishable as false. Furthermore, the attacker has access to process $w_n$ and measurement noise $z_n$. With this attack, the closed-loop systems above become $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{D}w_n + \underline{L}(z_n + v_n)$ and $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_n + \underline{\underline{D}}w_n + \underline{\underline{L}}(z_n + v_n)$.

### C. (Nonrobust) Dynamic Watermarking

The steady-state distribution of $\delta_n$ in an unattacked system will be Gaussian with mean-zero and a covariance matrix of

$$\Sigma_\Delta = (A + LC)\Sigma_\Delta(A + LC)^\mathsf{T} + \Sigma_W + L\Sigma_Z(\theta)L^\mathsf{T}. \tag{5}$$

Dynamic watermarking adds a small amount of Gaussian noise $e_n$, the values unknown to the attacker, into the control input $u_n = K\hat{x}_n + e_n$. This private excitation has mean-zero and covariance matrix $\Sigma_E$. Defining $\underline{B} = \begin{bmatrix} B^\mathsf{T} & B^\mathsf{T} \end{bmatrix}^\mathsf{T}$ and $\underline{\underline{B}} = \begin{bmatrix} B^\mathsf{T} & 0 \end{bmatrix}^\mathsf{T}$, the closed-loop systems with watermarking are given by $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{B}e_n + \underline{D}w_n + \underline{L}(z_n + v_n)$ and $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_t + \underline{\underline{B}}e_n + \underline{\underline{D}}w_n + \underline{\underline{L}}(z_n + v_n)$, respectively.

The watermarking noise $e_n$ leaves a detectable signal in the measurements $y_n$, which can detect the presence of an attack $v_n$ by comparing the observer error $C\hat{x}_n - y_n$ to previous values of the watermark $e_{n-k}$ for some integer $k > 0$. Specifically, the work in [26] proposes the tests

$$\begin{aligned} \text{as-}\lim_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} &= \\ C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z & \quad (6) \\ \text{as-}\lim_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)e_{n-k'-1}^\mathsf{T} = 0, & \quad (7) \end{aligned}$$

where $k' = \min_{k\geq 1}\{C(A + BK)^k B^\mathsf{T} \neq 0\}$. Any modeled attack passing these tests can be shown to asymptotically have zero power $\text{as-}\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} v_n^\mathsf{T} v_n = 0$ [26].

Finally, [26] also provides a test statistic for implementing the above test. Define $\psi_n = \begin{bmatrix}(C\hat{x}_n - y_n)^\mathsf{T} & e_{n-k'-1}^\mathsf{T}\end{bmatrix}^\mathsf{T}$ and $S_n = \sum_{i=n+1}^{n+\ell} \psi_n \psi_n^\mathsf{T}$. Then the negative log-likelihood of a Wishart distribution is

$$\begin{aligned}\mathcal{L} =& (m+q+1-\ell)\log\det S_n \\ &+ \operatorname{trace}\left\{ \begin{bmatrix}\left(C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z\right)^{-1} & 0 \\ 0 & \Sigma_E^{-1}\end{bmatrix} \times S_n \right\}.\end{aligned} \tag{DW}$$

This can be used to perform a statistical hypothesis test to detect attacks when using dynamic watermarking.

## III. Covariance-Robust Dynamic Watermarking

We develop covariance-robust dynamic watermarking methods for two different cases. The first is where $\theta$ is fixed but unknown, and the second is where $\theta$ is slowly varying.

### A. Fixed But Unknown Noise Covariance

We begin by stating our assumptions for this case. First, we assume that we have knowledge of a set of positive semidefinite matrices $\Sigma_{z,1}, \dots, \Sigma_{z,d}$ such that these matrices are affinely independent and $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$ for the set

$$\Omega^Z = \{\theta_1 \Sigma_{z,1} + \cdots + \theta_d \Sigma_{z,d} : \mathbf{1}^T\theta = 1, \theta \geq \mathbf{0}\}. \tag{8}$$

Note that $\Omega^Z$ is a polyhedron, and that this set is defined to be the convex combination of $\Sigma_{z,1}, \dots, \Sigma_{z,d}$. Our first result characterizes $\Omega^\Delta$, which is the set of possible $\Sigma_\Delta(\theta)$.

*Lemma 1:* Let $\bar{\Sigma}_{\delta,k}$ satisfy $\bar{\Sigma}_{\delta,k} = (A+LC)\bar{\Sigma}_{\delta,k}(A+LC)^T + \Sigma_W + L\Sigma_{z,k}L^T$. For $\Sigma_Z(\theta) = \theta_1\Sigma_{z,1} + \cdots + \theta_d\Sigma_{z,d}$, the solution to (5) is $\Sigma_\Delta(\theta) = \theta_1\bar{\Sigma}_{\delta,1} + \cdots + \theta_d\bar{\Sigma}_{\delta,d}$.

*Proof:* This immediately follows by noting that both sides of (5) are linear in the matrices $\Sigma_\Delta$ and $\Sigma_Z(\theta)$. ■

Since $E[\psi_n\psi_n^\mathsf{T}] = \operatorname{blkdiag}\{C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z, \Sigma_E\}$, we need to characterize the set $\Omega$ of feasible matrices in terms of $\theta$.

*Lemma 2:* Let $\bar{\Sigma}_k = \operatorname{blkdiag}\{C\bar{\Sigma}_{\delta,k}C^\mathsf{T} + \Sigma_{z,k}, \Sigma_E\}$. Then $\Omega = \{\theta_1\bar{\Sigma}_k + \cdots + \theta_d\bar{\Sigma}_d : \mathbf{1}^T\theta = 1, \theta \geq \mathbf{0}\}$.

*Proof:* This follows by the linearity in $\Sigma_\Delta$ and $\Sigma_Z$. ■

The set $\Omega$ represents covariance matrices of $\psi_n$ that are "acceptable", according to the original set $\Omega^Z$ of observation noise covariances that we should not mistake for attacks.

*Lemma 3:* The set $\Omega$ is of dimension $d-1$.

*Proof:* This follows from Lemma 1, the fact that $L$ is of full column-rank, and the observability of $(A+LC, C)$, which in turn follows from the observability of $(A, C)$. ■

Finally, consider a modification of (DW) given by

$$\begin{aligned}\mathcal{L}(S_n, V) = (m+q+1-\ell)\log\det S_n + \\ \operatorname{trace}\{VS_n\} - \ell\log\det V.\end{aligned} \tag{9}$$

Note (9) is the negative log-likelihood of an $(m+q)\times(m+q)$ Wishart distribution with scale matrix $V^{-1}$ and $\ell$ degrees of freedom. Now, we may present our test statistic. Let $\Omega^{-1} = \{V : V^{-1} \in \Omega\}$ and define the test statistic

$$T(S_n) = \min_{V\in\Omega^{-1}} \mathcal{L}(S_n, V) \tag{10}$$

for the composite null hypothesis $H_0 : E[\psi_n\psi_n^\mathsf{T}] \in \operatorname{int}(\Omega)$. For some $0 \leq \nu$, consider the test

$$\begin{cases}\text{reject } H_0 & \text{if } T(S_n) > \nu \\ \text{accept } H_0 & \text{if } T(S_n) \leq \nu.\end{cases} \tag{11}$$

Since $\arg\min_{V\in\Omega^{-1}} \mathcal{L}(S_n, V) = S_n^{-1}$, this proposed test is equivalent to the generalized likelihood ratio test.

*Theorem 1:* For large enough $\ell$, the decision rule (11) using test statistic $T(S_n)$ satisfies equal opportunity with respect to the null hypothesis and where the protected attribute is the true measurement noise covariance $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$.

*Proof:* Due to Lemma 3 and our assumption that $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$, $T(S_n)$ satisfies the Le Cam regularity conditions required for the application of Wilk's Theorem [27]. This means $-2T(S_n)$ will be asymptotically distributed as a $\chi^2(m+q-p)$ random variable plus a fixed constant *regardless of the true value of* $\Sigma_\Delta$, and thus implies that the event of a Type I error is independent of $\Sigma_\Delta$. ■

This is a useful result because it implies that, in the proper regime, our test can come arbitrarily close to satisfying the initial goal of remaining robust to some uncertainty in the distribution of the measurement noise. However, $\Omega^{-1}$ is a non-convex set, and so the computation of $T(S_n)$ is difficult. To this end, we propose the approximate test statistic

$$\begin{aligned}\bar{T}(S_n) = \min \quad & \mathcal{L}(S_n, V) \\ \text{s.t.} \quad & \begin{array}{rcl} \sum_{k=1}^p \theta_k \bar{\Sigma}_k^{-1} & \succeq & V, \\ \begin{bmatrix} V & I \\ I & \sum_{k=1}^p \theta_k\bar{\Sigma}_k\end{bmatrix} & \succeq & 0, \\ \mathbf{1}^\mathsf{T}\theta & = & 1, \\ \theta & \geq & \mathbf{0}. \end{array}\end{aligned} \tag{CRDW}$$

*Lemma 4:* For any $V \in \Omega^{-1}$, there exists a $\theta \in \mathbb{R}^p$ such that $(V, \theta)$ is a feasible solution to the optimization problem defining test (CRDW).

*Proof:* First observe that any $V \in \Omega^{-1}$ can be written as $V = (\sum_{k=1}^p \theta_k\bar{\Sigma}_k)^{-1}$ for some nonzero $\theta$ such that $\mathbf{1}^\mathsf{T}\theta = 1$. Thus, it holds trivially that

$$(\textstyle\sum_{k=1}^p \theta_k\bar{\Sigma}_k)^{-1} \succeq V \succeq (\sum_{k=1}^p \theta_k\bar{\Sigma}_k)^{-1} \tag{12}$$

The right-most constraint in (12) can be restated using the Schur complement, and this reformulation is exact. Since $\sum_{k=1}^p \theta_k\bar{\Sigma}_k \succeq 0$, the Schur complement implies the second constraint in (CRDW) is equivalent to $V - (\sum_{k=1}^p \theta_k\bar{\Sigma}_k)^{-1} \succeq 0$.

The first constraint in (CRDW) follows from the convexity of the matrix inverse for positive semidefinite matrices: Letting $X(\tau) = (1-\tau)X_1 + \tau X_2$ for positive definite $n\times n$ matrices $X_1, X_2$ and $0 \leq \tau \leq 1$, we have $\frac{\nabla^2}{\nabla\tau^2}X(\tau)^{-1} = 2X^{-1}(\tau)X'(\tau)X^{-1}(\tau)X'(\tau)X^{-1}(\tau)$. For any $a \in \mathbb{R}^n$, the function $\phi_a(\tau) = a^\mathsf{T}X^{-1}(\tau)a$ will have second derivative $\phi_a''(\tau) = 2a^\mathsf{T}X^{-1}(\tau)X'(\tau)X^{-1}(\tau)X'(\tau)X^{-1}(\tau)a \geq 0$ due to the positive-semidefiniteness of $X(\tau)^{-1}$, so $(1-\tau)\phi_a(0) + \tau\phi_a(1) \geq \phi_a(\tau)$. Since this holds for any $a$, we have that

$$\textstyle\sum_{k=1}^p \theta_k\bar{\Sigma}_k^{-1} \succeq (\sum_{k=1}^p \theta_k\bar{\Sigma}_k)^{-1}. \tag{13}$$

The first constraint in (CRDW) follows from (12) and (13). ■

*Remark 1:* It was shown in [26] that test (7) ensures $\alpha = 0$ in any attack such that it holds true. In that case, we have

$$\begin{aligned} &\text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} \\ &\quad = C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta) + \\ &\Sigma_S + \text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T} C^\mathsf{T}, \end{aligned} \quad (14)$$

since the Schur stability of $A + BK$ implies that any effect of $x_0$ and $\eta_0$ are reduced to zero asymptotically. Since $\Sigma_S$ and $\text{as-lim}_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T} C^\mathsf{T}$ are both positive semidefinite, meaning that

$$\begin{aligned} &\text{as-lim} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} \succeq \\ &\qquad C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta). \end{aligned} \quad (15)$$

Inverting both sides of this implies that, in the case that $\Sigma_S + \text{as-lim}_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T} C^\mathsf{T} \neq 0$, we can generally expect that $S_n^{-1} \preceq (C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta))^{-1} \in \Omega^{-1}$. The takeaway is that the looseness of the upper bound (13) should not greatly decrease the power of the modified test in the presence of test (7), as the tight lower bound is more germane to situations where the system is actually being attacked.

*Remark 2:* If the dimension $m + q$ is large, then the optimization (CRDW) may be expensive to solve from scratch each time. Furthermore, $S_n$ will likely not change drastically between runs when $\ell$ is large. So, lighter-weight first-order methods such as ADMM can be used instead [28]. These generally take longer to converge to high levels of accuracy, but have the advantage of being able to be readily warm-started.

### *B. Slowly Varying Unknown Noise Covariance*

A key difference between this setting and that of the static distribution is that a shift in the observer noise covariance in one period can have impacts on $\Sigma_\Delta$ over the next few periods that do not easily fit into our previous representation of the $\Omega$. This is because it will take many steps before the covariance of $\delta_n$ approaches its asymptotic limit in $\Omega$. Thus, to accommodate a dynamically changing distribution of $z_n$, we must use an expansion of the set $\Omega$.

We modify our setup for this subsection. The true covariance of $\delta_n$ and $z_n$ are $\Sigma_{\Delta_n}$ and $\Sigma_{Z_n}$, respectively. Let $\Psi_n = \Sigma_{Z_n} - \Sigma_{Z_{n-1}}$ and $\Phi_n^j = (A + LC)^j L \Psi_n L^\mathsf{T} (A + LC)^{j^\mathsf{T}}$. Note that all $\Sigma_{Z_n}$ are still assumed to be in $\Omega^Z$. Finally, we make some additional assumptions. Since the spectral radius of $A + LC$ is less than one, there exists some induced norm (denote this $\|\cdot\|$) such that $\|A + LC\| < 1$ [29]. We assume $\theta$ changes every step but $\Sigma_{Z_0} \in \Omega$ and all $\Psi_n$ satisfy $\|\Psi_n\| \leq \xi$ for some known value of $\xi > 0$. We also assume the system starts at steady state in the sense $\Sigma_{\Delta_0} = (A + LC)\Sigma_{\Delta_0}(A + LC)^\mathsf{T} + \Sigma_W + L\Sigma_{Z_0}L^\mathsf{T}$. Under these assumptions we have:

*Lemma 5:* Let $\varepsilon \in \mathbb{R}$ be defined as

$$\varepsilon = \frac{\xi \|C\|^2 \|L\|^2 \|A + LC\|^2 \sqrt{m}}{(1 - \|A + LC\|^2)^2} \quad (16)$$

Then $C\Sigma_{\Delta_n} C^\mathsf{T} + \Sigma_{Z_n} \in \Omega \oplus \{E : -\varepsilon I \preceq E \preceq \varepsilon I\}$, where $\oplus$ is the Minkowski sum for all $n$.

*Proof:* Let $\Omega_{m\times m}$ be the set of $m \times m$ upper-left submatrices of elements of $\Omega$, associated with $C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta)$ terms. We start by noting that

$$\begin{aligned} \Sigma_{\Delta_1} =& (A + LC)\Sigma_{\Delta_0}(A + LC)^\mathsf{T} + \Sigma_W + L\Sigma_{Z_1}L^\mathsf{T} \\ =& \Sigma_{\Delta_0} + \Phi_1^0. \end{aligned} \quad (17)$$

Similarly, we can see that $\Sigma_{\Delta_2} = \Sigma_{\Delta_0} + L\left(\Psi_0 + \Psi_1\right)L^\mathsf{T} + \Phi_0^1 = \Sigma_{\Delta_0} + \Phi_2^0 + \Phi_1^0 + \Phi_1^1$. Continuing this recursion relation leads to the fact that

$$\Sigma_{\Delta_n} = \Sigma_{\Delta_0} + \textstyle\sum_{i=0}^{n-1} \sum_{j=0}^{i} \Phi_{n-i}^j. \quad (18)$$

Due to the Schur stability of $A + LC$, the following limit exists, and can be represented as in Lemma 1.

$$\Sigma_{\Delta_\infty^{k'}} = \lim_{k\to\infty} \left(\Sigma_{\Delta_0} + \textstyle\sum_{i=k-k'}^{k-1} \sum_{j=0}^{i} \Phi_{k-i}^j\right) \quad (19)$$

Note that $\Sigma_{\Delta_\infty^{k'}}$ is the steady state that $\Sigma_{\Delta_n}$ would ultimately reach if $\theta$ (and therefore $\Sigma_{Z_n}$ does not shift after step $k'$; thus, it solves (5) for $\Sigma_{Z_{k'}}$ and exists in $\Omega_{m\times m}$. Denote $\Upsilon_i = \Sigma_{\Delta_\infty^i} - \Sigma_{\Delta_\infty^{i-1}}$. Then,

$$\begin{aligned} \Sigma_{\Delta_n} &= \lim_{k\to\infty} \left(\Sigma_{\Delta_0} + \textstyle\sum_{i=k-n}^{k-1} \sum_{j=0}^{i-k+n} \Phi_{k-i}^j\right) \\ &= \Sigma_{\Delta_\infty^n} - \lim_{k\to\infty} \left(\textstyle\sum_{i=k-n}^{k-1} \left(\sum_{j=i-k+n+1}^{i} \Phi_{k-i}^j\right)\right) \\ &= \Sigma_{\Delta_\infty^n} - \textstyle\sum_{i=1}^{n} (A + LC)^{n-i+1} \Upsilon_i (A + LC)^{n-i+1^\mathsf{T}} \end{aligned} \quad (20)$$

Note that the term in the limit in the first equality is a constant in $k$ due to a simple re-indexing of (18). This is convenient because we can now break $\Sigma_{\Delta_n}$ into an element known to be in $\Omega_{m\times m}$ and an error term. Our goal is now to choose $\varepsilon$ large enough to bound

$$\min_{P\in\Omega_{m\times m}} \left\|C\Sigma_{\Delta_n}C^\mathsf{T} + \Sigma_{Z_n} - P\right\|_2, \quad (21)$$

over all paths that $\Sigma_{Z_n}$ can take. An easy bound on the minimization is to simply set $P = C\Sigma_{\Delta_\infty^n}C^\mathsf{T} + \Sigma_{Z_n}$. Then, $\varepsilon$ only needs to exceed

$$\left\|\textstyle\sum_{i=1}^{n} C(A + LC)^{n-i+1} \Upsilon_i (A + LC)^{n-i+1^\mathsf{T}} C^\mathsf{T}\right\|_2 \quad (22)$$

By sub-multiplicativity of induced norms,

$$\begin{aligned} \|\Upsilon_i\| &= \left\|\textstyle\sum_{j=0}^{\infty} \Phi_i^j\right\| \leq \textstyle\sum_{j=0}^{\infty} \|(A + LC)\|^{2j} \|L\| \|\Psi_{n'+k_i}\| \\ &= \xi \|L\|^2 \left(1 - \|A + LC\|^2\right)^{-1} \end{aligned} \quad (23)$$

Finally, using the fact that $\|\cdot\|_2 \leq \sqrt{m}\|\cdot\|$ [30] and applying (23) to the error term from (21) yields the desired result. ■

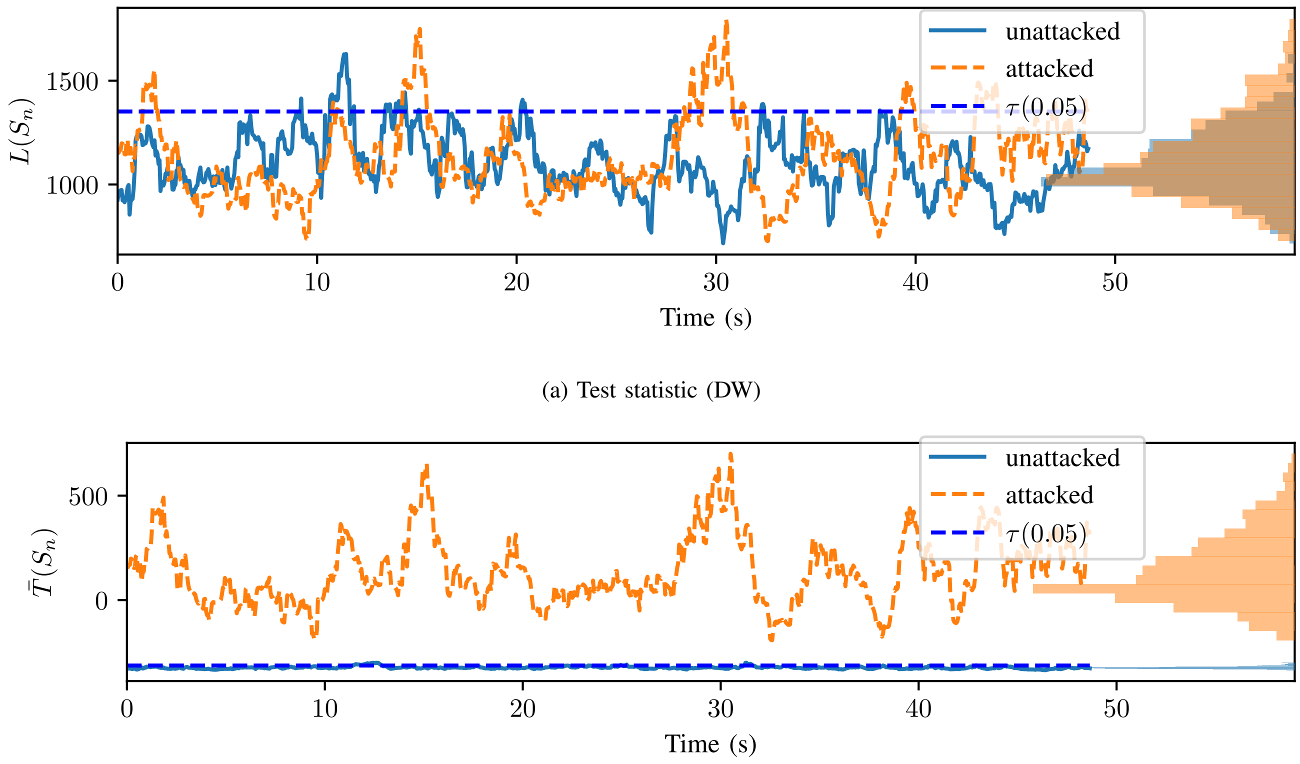
*Remark 3:* Due to the topological equivalence of induced norms, the dependence of our choice of norm $\|\cdot\|$ on $A + LC$ can only affect the value of $\xi$ required by a constant $\sqrt{m}$.

*Corollary 1:* If $\|A + LC\|_2 < 1$, then the statement in Lemma 5 holds for $\|\cdot\| = \|\cdot\|_2$ and

$$\varepsilon = \frac{\xi \|C\|_2^2 \|L\|_2^2 \|A + LC\|_2^2}{(1 - \|A + LC\|_2^2)^2} \quad (24)$$

*Proof:* The proof of this result is almost identical to the proof of the previous lemma with the only changes that $\|\cdot\| = \|\cdot\|_2$ and that we stop after applying (23) to (21). ■

With this $\varepsilon$, it is straightforward to extend the previous test statistic (CRDW) to this new expansion of $\Omega$ as long as



(a) Test statistic (DW)

(b) Test statistic (CRDW)

Fig. 1: The evolution and histogram of test statistics (DW) and (CRDW) on the attacked and unattacked systems where $\Sigma_Z$ is fixed, but unknown to the tester. In this case, the nonrobust test statistic (DW) is unable to clearly distinguish the attacked from the unattacked system, whereas the new test statistic (CRDW) can.

$\bar{\Sigma}_k - \varepsilon I$ remains positive definite for all $k$. In this case, we may define our new test statistic as

$$\begin{aligned} \underline{T}(S_n) = \min \quad & \mathcal{L}(S_n, V) \\ \text{s.t.} \quad & \textstyle\sum_{k=1}^{p} \theta_k \left(\bar{\Sigma}_k - \varepsilon I\right)^{-1} && \succeq && V, \\ & \begin{bmatrix} V & I \\ I & \varepsilon I + \sum_{k=1}^{p} \theta_k \bar{\Sigma}_k \end{bmatrix} && \succeq && 0, \\ & \mathbf{1}^\mathsf{T}\theta && = && 1, \\ & \theta && \geq && \mathbf{0}. \end{aligned} \tag{CRDW*}$$

*Remark 4:* If there is some $k$ so $\bar{\Sigma}_k - \varepsilon I$ is not positive definite, then the first constraint above is not well-defined. Recalling that $V$ is a surrogate for $\left(C\Sigma_{\Delta_n}C^\mathsf{T} + \Sigma_{Z_n}\right)^{-1}$, we note $V$ trivially satisfies $\Sigma_{Z_n}^{-1} \succeq V$. Thus in this problematic case, we may replace the $\left(\bar{\Sigma}_k - \varepsilon I\right)$ in the first constraint with $\Sigma_{z,k}$, for all $k$. This issue is unlikely to be of practical concern for the same reasons discussed in Remark 1 regarding the relaxation of the set $\Omega$. Specifically, the structure of the attacks makes it unlikely that the first constraint in (CRDW*) would be binding in any case.

## IV. Empirical Results

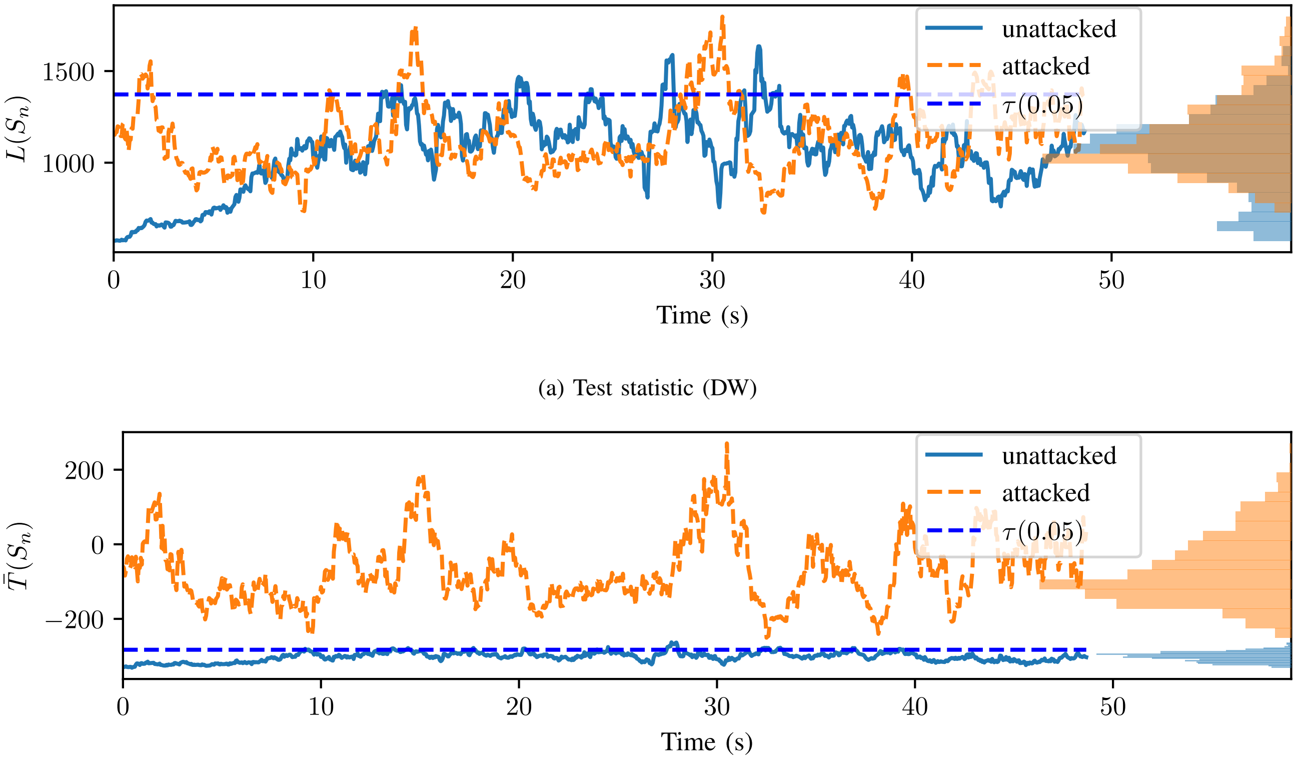
In this section, we present simulation results that showcase the strength of our method when compared with the original test statistic (DW). We present results for both the case where the noise distribution is fixed but unknown, and for the case where the noise covariance is unknown and slowly-varying.

We use the standard model for simulation of an autonomous vehicle in [31], where the error kinematics of lane keeping and speed control is given by $x^\mathsf{T} = \begin{bmatrix} \psi & y & s & \gamma & v \end{bmatrix}$ and $u^\mathsf{T} = \begin{bmatrix} r & a \end{bmatrix}$. Here, $\psi$ is heading error, $y$ is lateral error, $s$ is trajectory distance, $\gamma$ is vehicle angle, $v$ is vehicle velocity, $r$ is steering, and $a$ is acceleration. We linearize and initialize with a straight trajectory and constant velocity $v_0 = 10$. We then performed exact discretization with sampling period $t_s = 0.05$. This yields the system dynamics

$$A = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{10} & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{40} & 0 \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{400} & 0 \\ \frac{1}{2400} & 0 \\ 0 & \frac{1}{800} \\ \frac{1}{20} & 0 \\ 0 & \frac{1}{20} \end{bmatrix} \tag{25}$$

with $C = \begin{bmatrix} I & 0 \end{bmatrix} \in \mathbb{R}^{3\times 5}$. We use process noise covariance $\Sigma_W = 10^{-8} \times I$.

All tests use dynamic watermarking with variance $\Sigma_E = \frac{1}{2}I$, and $K$ and $L$ were chosen to stabilize the system without an attack. We conduct four simulations: attacked and non-attacked systems where the measurement noise covariance



(a) Test statistic (DW)

(b) Test statistic (CRDW*)

Fig. 2: The evolution and histogram of test statistics (DW) and (CRDW*) on the attacked and unattacked systems where $\Sigma_Z$ varies as described, again unknown to the tester. Note the robust statistic (CRDW*) takes distinctly higher for the attacked values over almost the entire 1000 iterations than in the unattacked system, while the nonrobust statistic (DW) is again unable to clearly distinguish the two.

is fixed, and attacked and non-attacked systems where the measurement noise covariance is allowed to vary. We ran all four simulations for 1000 iterations, or 50 seconds. In all cases, we compare the test metrics using the hypothesis test described in (11), where the measurement noise covariance is assumed to be $10^{-5} \times I$. When simulating the attacked system, we choose an attacker with $\alpha = -1$, $\eta_0 = 0$, $\Sigma_O = 10^{-8} \times I$, and $\Sigma_S = 10^{-8} \times I$.

### A. *Fixed Covariance*

We first show our test outperforms in the case where the true measurement noise covariance matrix is fixed but unknown to the tester. In our simulations, the true noise covariance is $\Sigma_Z = 10^{-5} \times \text{diag}\{0.18, 30, 0.18\}$. In all tests, $\Omega^Z$ is described by the $p = 4$ extreme points: $\Sigma_{Z,1} = 10^{-6} \times \text{diag}\{300, 1.8, 1.8\}$, $\Sigma_{Z,2} = 10^{-6} \times \text{diag}\{1.8, 300, 1.8\}$, $\Sigma_{Z,3} = 10^{-6} \times \text{diag}\{9, 9, 12\}$, $\Sigma_{Z,4} = 10^{-6} \times \text{diag}\{9, 9, 9\}$. Both the true measurement noise covariance and that incorrectly assumed in test statistic (DW) are in the resulting set. The simulation is run for 1000 steps.

Fig. Figure 1 shows the efficacy of our method under this new uncertainty. If test detection is consistent, the negative log likelihood values should be lower under regular conditions, and higher when the model is attacked. In particular, the nonrobust test statistic (DW) is shown in Fig. 1a to be wholly unable to distinguish an attacked system from an unattacked system when its assumption on the measurement noise covariation is violated, while Fig. 1b shows the robust test statistic (CRDW) to be able to do so.

### B. *Varying Covariance*

Unattacked and attacked simulations were also conducted with a measurement noise distribution that was allowed to vary. We set $\tau = 1$ and $\xi = 0.00002$, implying $\varepsilon = 7.205 \times 10^{-6}$. The true measurement noise is initialized at $\Sigma_{Z_0} = 10^{-5} \times \text{diag}\{0.9, 0.9, 1.2\}$. This shifts linearly over the course of 250 iterations to a new value of $\Sigma_{Z_{250}} = 10^{-5} \times \text{diag}\{15, 15, 0.18\}$, at which point it changes direction to shift linearly over 250 iterations to a value of $\Sigma_{Z_{500}} = 10^{-5} \times \text{diag}\{30, 0.18, 0.18\}$. The measurement noise covariance stays at this value for 150 iterations. It then shifts linearly over 200 iterations to a terminal value of $\Sigma_{Z_{850}} = 10^{-5} \times \text{diag}\{0.18, 30, 0.18\}$, which it takes for another 150 iterations before the simulation is terminated. The results for both the nonrobust and robust tests are shown in Fig. 2. As in the fixed covariance case, our test is able to distinguish between the attacked and unattacked systems better and more consistently than the nonrobust test that requires unsatisfied assumptions.

## V. Conclusion

We developed covariance-robust dynamic watermarking tests for detecting sensor attacks on LTI systems in the presence of uncertainty about the measurement noise covariance. We considered cases where the covariance of measurement noise is unknown and either fixed or slowly-varying, and we required our test to be "fair" with respect to all possible values of the covariance in that it not be more or less powerful for some covariances over others. These reflect real-world needs that will increase as 5G is deployed, because there will be an increase in the deployment of smart CPS systems. In such systems, an "unfair" test can translate to disparate impact across different users in different environments, which is a problem of algorithmic bias. Future research includes studying how dynamic watermarking can be adapted to other system uncertainties.

## References

[1] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Security for 4g and 5g cellular networks: A survey of existing authentication and privacy-preserving schemes," *Journal of Network and Computer Applications*, vol. 101, pp. 55–82, 2018.

[2] M. Abrams and J. Weiss, "Malicious control system cyber security attack case study–Maroochy water services, australia," *MITRE*, 2008.

[3] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[4] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[5] B. Satchidanandan and P. Kumar, "Dynamic watermarking: Active defense of networked cyber-physical systems," *Proc. of IEEE*, 2016.

[6] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *Proc. of IEEE CDC*, 2014, pp. 3757–3764.

[7] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE CST*, vol. 22, no. 4, pp. 1396–1407, 2014.

[8] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5g systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[9] I. Ahmad, Z. Kaleem, R. Narmeen, L. D. Nguyen, and D.-B. Ha, "Quality-of-service aware game theory-based uplink power control for 5g heterogeneous networks," *Mobile Networks and Applications*, vol. 24, no. 2, pp. 556–563, 2019.

[10] B. Satchidanandan and P. Kumar, "On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems," in *IEEE COMSNETS*, 2017, pp. 23–30.

[11] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conference*. IEEE, 2009, pp. 911–918.

[12] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *Proc. of IEEE CDC*, 2010, pp. 5967–5972.

[13] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, 2015.

[14] M. Porter, S. Dey, A. Joshi, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, "Detecting deception attacks on autonomous vehicles via linear time-varying dynamic watermarking," *arXiv preprint arXiv:2001.09859*, 2020.

[15] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: the state of the art," *arXiv preprint arXiv:1703.09207*, 2017.

[16] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, 2016, pp. 3315–3323.

[17] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *IEEE ICDMW*, 2009, pp. 13–18.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 2012, pp. 214–226.

[19] I. Zliobaite, "On the relation between accuracy and fairness in binary classification," *arXiv preprint arXiv:1505.05723*, 2015.

[20] M. Olfat and A. Aswani, "Spectral algorithms for computing fair support vector machines," in *AISTATS*, 2018, pp. 1933–1942.

[21] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *arXiv preprint arXiv:1703.00056*, 2017.

[22] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017.

[23] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," in *NeurIPS*, 2017, pp. 5036–5044.

[24] M. Olfat and A. Aswani, "Convex formulations for fair principal component analysis," in *AAAI*, vol. 33, 2019, pp. 663–670.

[25] A. Aswani and M. Olfat, "Optimization hierarchy for fair statistical decision problems," *arXiv preprint arXiv:1910.08520*, 2019.

[26] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general LTI systems," in *IEEE CDC*, 2017, pp. 1834–1839.

[27] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The annals of mathematical statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[28] Z. Wen, D. Goldfarb, and W. Yin, "Alternating direction augmented lagrangian methods for semidefinite programming," *Mathematical Programming Computation*, vol. 2, no. 3-4, pp. 203–230, 2010.

[29] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[30] B. Q. Feng, "Equivalence constants for certain matrix norms," *Linear algebra and its applications*, vol. 374, pp. 247–253, 2003.

[31] V. Turri, A. Carvalho, H. Tseng, K. Johansson, and F. Borrelli, "Linear model predictive control for lane keeping and obstacle avoidance on low curvature roads," in *Proc. of IEEE ITSC*, 2013, pp. 378–383.