RoS-Guard: Robust and Scalable Online Change Detection with Delay-Optimal Guarantees

Zelin Zhu[1*], **Yancheng Huang**[1*], **Kai Yang**[1,2,3†]
[1]School of Computer Science and Technology, Tongji University, China.
[2] Shenzhen Loop Area Institute, China.
[3] MOE Key Laboratory of Embedded Systems and Service Computing of Tongji University, China
zelinzhu@tongji.edu.cn, 2130783@tongji.edu.cn, kaiyang@tongji.edu.cn

Abstract

Online change detection (OCD) aims to rapidly identify change points in streaming data and is critical in applications such as power system monitoring, wireless network sensing, and financial anomaly detection. Existing OCD methods typically assume precise system knowledge, which is unrealistic due to estimation errors and environmental variations. Moreover, existing OCD methods often struggle with efficiency in large-scale systems. To overcome these challenges, we propose RoS-Guard, a robust and optimal OCD algorithm tailored for linear systems with uncertainty. Through a tight relaxation and reformulation of the OCD optimization problem, RoS-Guard employs neural unrolling to enable efficient parallel computation via GPU acceleration. The algorithm provides theoretical guarantees on performance, including expected false alarm rate and worst-case average detection delay. Extensive experiments validate the effectiveness of RoS-Guard and demonstrate significant computational speedup in large-scale system scenarios.

Extended version — https://github.com/RosGuard-Extended/Extended-version

Introduction

Online Change Detection (OCD) aims to detect distributional changes in a data stream as new observations arrive sequentially, often balancing detection delay and false alarm rate. The OCD problem in dynamic systems arises in various fields, including detecting false data injection attacks (FDIAs) in smart grids and identifying anomalies in wireless networks, as well as other information technology (IT) systems (Yang et al. 2016; Dou, Yang, and Poor 2019).

Compared to OCD problems under stationarity assumptions, OCD in dynamic systems is particularly challenging because the inherent dynamics of the system can induce significant distributional changes over time, even in the absence of external events. These time-varying characteristics of pre- and post-change distributions complicate the task of reliably identifying abrupt changes. Existing studies (Zhang and Wang 2021; Huang et al. 2011; Li, Yılmaz, and Wang 2014) often presuppose precise knowledge of the system. However, in practical scenarios, the parameters of the system commonly exhibit change-unrelated uncertainties due to both imperfect information acquisition and inherent dynamics. For example, in smart grids, environmental fluctuations can alter line admittances, leading to inaccuracies in the system matrix and reducing the effectiveness of false data injection attack detection (Li, Yılmaz, and Wang 2014). Similarly, in MIMO systems, aging, quantization, and estimation errors often hinder accurate channel matrix estimation (Weber, Sklavos, and Meurer 2006). These challenges highlight the need for a robust detection framework with solid theoretical support.

The performance of OCD methods depends not only on the delay introduced by the detection procedure itself, but also on the computational efficiency of the algorithm, as its execution time may span a significant number of data arrivals, especially when the system dimension scales up, thereby affecting how promptly a change can be declared. Unlike offline methods (Yu et al. 2024; Li, Cao, and Meng 2024), which benefit from pre-collected data and parallel computation, OCD must process streaming data under real-time constraints, making efficiency improvements significantly more challenging and heavily dependent on algorithmic innovations.

To improve OCD efficiency, recent works have explored approaches such as buffering a fixed number of m steps of observations (Zhang et al. 2024) to facilitate efficient change detection in hidden Markov models, and bandit-based selective sensing (Gopalan, Lakshminarayanan, and Saligrama 2021) to reduce observation costs by querying only a subset of informative sensors at each step. However, these methods either compromise real-time responsiveness or involve partial observability, and do not address model uncertainty. While (Zhang and Wang 2021) considers unknown, time-varying changes and achieves linear complexity in the number of observations, but its scalability with respect to system dimension remains unexamined.

Overall, the limited research on OCD methods under system uncertainty, combined with the significant challenge of improving detection efficiency in strict online settings, motivated us to design a robust and computationally efficient OCD algorithm that also scales effectively with system di-

[*]These authors contributed equally. Author order determined by coin flip.
[†]Corresponding author

mensionality. The key contributions of this paper are summarized as follows:

- We propose RoS-Guard, a robust OCD algorithm that accounts for system uncertainty with theoretical guarantees.
- To improve detection efficiency and ensure scalability in large-scale systems, a method leveraging neural unrolling and GPU acceleration is introduced.
- Theoretical analysis of RoS-Guard's performance is presented, and extensive experiments demonstrate the effectiveness and acceleration of the proposed method in large-scale settings.

Related Work

Early approaches to online change detection (OCD) often assume complete system knowledge and static pre-/post-change distributions. Classical methods such as those in (Huang et al. 2011; Li, Yılmaz, and Wang 2014) adopt CUSUM-type statistics under known dynamics, with (Li, Yılmaz, and Wang 2014) further assuming fixed change locations—limiting applicability in dynamic settings. To handle model uncertainty, (Unnikrishnan, Veeravalli, and Meyn 2011) studies robust OCD using least favorable distributions (LFDs) within known uncertainty sets, while (Molloy and Ford 2017) relaxes the assumptions, offering theoretical guarantees for misspecified detection rules. Other works (Huang et al. 2014; Zhang and Wang 2021) rely on residual-based detection, but often face challenges such as ill-conditioned covariances or relaxed formulations that hurt detection quality.

More recently, (Hare, Kaplan, and Veeravalli 2021) introduced the Uncertain Likelihood Ratio (ULR) test to handle partially known distributions, followed by (Hare and Kaplan 2022), a faster variant. (Xie 2022) adopts a nonparametric Wasserstein ambiguity set to improve robustness. While these methods advance uncertainty handling, they often make simplifying assumptions: known ambiguity sets, independent or Markovian dynamics, fixed detection structure, or low-dimensional observations.

Emerging studies also consider computational efficiency of OCD algorithms. For example, (Gopalan, Lakshminarayanan, and Saligrama 2021) develops bandit-style methods with partial observations, and (Zhang et al. 2024) accelerates detection in hidden Markov models via buffered schemes. Yet these methods, they do not consider uncertainties inherent in the system parameters, which is critical for robust detection in practical dynamic environments. Unlike offline methods that exploit data parallelism, online change detection relies heavily on fast optimization to ensure real-time performance. Neural unrolling has been shown effective in accelerating large-scale constrained optimization with GPU support (Shi et al. 2021; He et al. 2022; Chen and Yang 2025).

Based on the existing research, we consider online change detection under more general system uncertainty conditions. Inspired by neural unrolling, we focus on designing scalable algorithms that efficiently handle large-scale and complex system settings.

Problem Statement

We consider the online change detection (OCD) problem in uncertain dynamic systems. The goal is to detect distributional changes in a sequence of observations as quickly as possible while satisfying a false alarm constraint. Our modeling follows Lorden's minimax optimality criterion for worst-case detection delay.

Let Γ denote the stopping time when a change is declared. The detection performance is measured by Lorden's worst-case Average Detection Delay (ADD) (Lorden 1971):

$$J(\Gamma) \triangleq \sup_{\tau} \operatorname*{ess\,sup}_{\mathcal{A}\tau} \mathbb{E}\tau \left[(\Gamma - \tau)^+ \mid \mathcal{A}\tau \right], \tag{1}$$

where $\mathcal{A}\tau$ is the σ-algebra generated by observations up to time τ, and $\mathbb{E}\tau$ is the conditional expectation given a change occurs at time τ. To control false alarms, the expected run length under no change, $\mathbb{E}\infty[\Gamma]$, must exceed a threshold $\beta > 0$. Thus, the OCD problem is formulated as:

$$\inf_{\Gamma} J(\Gamma) \quad \text{subject to} \quad \mathbb{E}_\infty[\Gamma] \geq \beta. \tag{2}$$

We focus on systems where the observation at time t follows a dynamic linear model with time-varying parameters:

$$\mathbf{x}^{(t)} = \begin{cases} \mathbf{H}\boldsymbol{\theta}^{(t)} + \mathbf{n}^{(t)}, & t < t_a, \\ \left(\mathbf{H} + \Delta\mathbf{H}^{(t)}\right)\boldsymbol{\theta}^{(t)} + \mathbf{n}^{(t)}, & t \geq t_a, \end{cases} \tag{3}$$

where $\mathbf{x}^{(t)} \in \mathbb{R}^M$ is the observed signal, $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^N$ is an unknown system state, and $\mathbf{H} \in \mathbb{R}^{M\times N}$ is the nominal system matrix. The noise $\mathbf{n}^{(t)}$ is i.i.d. Gaussian, i.e., $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_M)$. The post-change system is affected by unknown perturbations $\Delta\mathbf{H}^{(t)}$ from time t_a onward.

Unlike previous works, we recognize that the precise knowledge of system matrix is typically unattainable in practical scenarios, and therefore assume it belongs to an uncertainty set $\mathcal{S}$, i.e., $\mathbf{H} \in \mathcal{S}$. Let change in observations as $\mathbf{a}^{(t)} \triangleq \Delta\mathbf{H}^{(t)}\boldsymbol{\theta}^{(t)}$. The goal of the OCD is to detect the injected vector $\mathbf{a}^{(t)}$ as soon as possible after it actually occurs at time instant t_a.

Model (3) can be naturally generalized to long-range temporal systems by allowing the hidden state $\boldsymbol{\theta}^{(t)}$ to encode information from multiple past time steps, and by scaling up the dimensions of both $\mathbf{H}$ and $\boldsymbol{\theta}^{(t)}$ to capture richer dependencies and longer temporal horizons. It also captures a broad class of uncertain linear systems where both the state vector and observation model evolve over time, and the change manifests as a structured perturbation to the system matrix (e.g. wireless MIMO systems(Pei et al. 2011; Ghavami and Naraghi-Pour 2017)), smart grid systems(Li and Wang 2015)).

RoS-Guard

This section introduces RoS-Guard, focusing on the problem modeled in (3). Specifically, (i) we begin with the reformulation of the Generalized Log-Likelihood Ratio (GLLR), which serves as the statistical evidence for detecting changes; (ii) the estimation of GLLR with system uncertainty is then formulated as a mixed-integer quadratic programming (MIQP) problem; (iii) to address computational

challenges in high-dimensional settings, a relaxation and decomposition strategy is applied; and (iv) neural unrolling is employed to develop a scalable approximate solver that leverages GPU acceleration for fast computation. Finally, the overall workflow of RoS-Guard is summarized.

Generalized Log-likelihood Ratio

From the model (3), the unknown and time-varying nature of $\boldsymbol{\theta}^{(t)}$ makes the component of $\mathbf{a}^{(t)}$ within the column space of $\mathbf{H}$ intrinsically unobservable. To address this, we extract the component of $\mathbf{a}^{(t)}$ that lies in the orthogonal complement of $\mathcal{C}(\mathbf{H})$ by introducing

$$\boldsymbol{\mu}^{(t)} \triangleq \mathbf{P}_{\mathbf{H}}^{\perp}\mathbf{a}^{(t)} \in \mathcal{C}^{\perp}(\mathbf{H}), \tag{4}$$

where $\mathbf{P}_{\mathbf{H}}^{\perp} \triangleq \mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is the projection matrix onto $\mathcal{C}^{\perp}(\mathbf{H})$. To ensure the reliability of change detection, we assume a bounded magnitude for the nonzero entries of $\boldsymbol{\mu}^{(t)}$ i.e. $\rho_L \leq |\mu_m^{(t)}| \leq \rho_H, m \in \mathcal{U}^{(t)}$, where $\mathcal{U}^{(t)}$ denotes the support of $\boldsymbol{\mu}^{(t)}$. The lower bound ρ_L filters out noise-induced small perturbations, while the upper bound ρ_H guards against extreme outliers due to model mismatch or measurement errors.

By estimating $\boldsymbol{\theta}^{(t)}$ and $\Delta\mathbf{H}^{(t)}$ via maximum likelihood (MLE) (Tartakovsky, Nikiforov, and Basseville 2014a), the detection rule follows the generalized likelihood ratio (GLR) form:

$$\Gamma_R = \min\left\{K : \max_{1\leq k\leq K} \Lambda_k^{(K)} \geq h\right\}, \tag{5}$$

where $\Lambda_k^{(K)}$ denotes the log-likelihood ratio statistic between the pre- and post-change models.

Incorporating the orthogonal component $\boldsymbol{\mu}^{(t)}$ defined in (4), the post-change model can be reformulated to depend only on the detectable directions. As a result, the log-likelihood ratio becomes:

$$\Lambda_k^{(K)} \triangleq \sup_{\{\mathcal{U}^{(t)}\},\mathbf{H}} \ln \frac{\sup_{\boldsymbol{\theta}^{(t)},\Delta\mathbf{H}^{(t)},\boldsymbol{\mu}^{(t)}} \prod_{t=k}^{K} f_q\left(\mathbf{x}^{(t)} \mid \boldsymbol{\theta}^{(t)}, \mathbf{H}, \Delta\mathbf{H}^{(t)}\right)}{\sup_{\boldsymbol{\theta}^{(t)}} \prod_{t=k}^{K} f_p\left(\mathbf{x}^{(t)} \mid \boldsymbol{\theta}^{(t)}, \mathbf{H}\right)}, \tag{6}$$

where f_p and f_q are Gaussian densities corresponding to the pre- and post-change models, respectively. $\Lambda_k^{(K)}$ can be further simplified to a sum of scalar statistics, i.e., $\Lambda_k^{(K)} = \sum_{t=k}^{K} v_t$, where each v_t represents the instantaneous evidence for change, and is derived as follows:

$$\begin{aligned} v_t = \sup_{\mathcal{U}^{(t)}} \sup_{\boldsymbol{\mu}^{(t)},\mathbf{H}} \quad & \frac{1}{2\sigma^2}\left\{2\left(\boldsymbol{\mu}^{(t)}\right)^T \boldsymbol{x}^{(t)} - \left\|\boldsymbol{\mu}^{(t)}\right\|_2^2\right\} \\ \text{s.t.} \quad & \rho_L \leq \left|\mu_m^{(t)}\right| \leq \rho_U, \quad \forall m \in \mathcal{U}^{(t)}, \\ & \mathbf{H}^T\boldsymbol{\mu}^{(t)} = \mathbf{0}, \\ & \mathbf{H} \in \mathcal{S}. \end{aligned} \tag{7}$$

Recursively, the accumulated evidence statistic V_K can be computed as

$$V_K \triangleq \max_{1\leq k\leq K} \Lambda_k^{(K)} = \max\{V_{K-1}, 0\} + v_K, \tag{8}$$

where $V_0 = 0$ and a change is declared when $V_K \geq h$.

Full derivation of $\Lambda_k^{(K)}$ is provided in the extended version (see the Links section)

GLLR with System Uncertainty

In this subsection, we focus on the uncertainty of the system matrix and detail the reformulation of equation (7) as an MIQP problem. To simplify notation, we omit the time superscripts of variables.

We first reformulate the original sup-based objective v_t as an equivalent inf minimization problem for tractability and we represent the support set of the decision vector $\boldsymbol{\mu}$ using a binary vector $\boldsymbol{u} \in \{0,1\}^N$, where $u_m = 1$ indicates that the m-th component of $\boldsymbol{\mu}$ belongs to the active support set $\mathcal{U}$. Subsequently, we decomposed $\boldsymbol{\mu}$ into two nonnegative components with complementarity constraint, i.e, $\boldsymbol{\mu} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-, {\boldsymbol{\mu}^+}^T\boldsymbol{\mu}^- = \mathbf{0}$.

Then we employ the versatile constraint-wise uncertainty paradigm (Yang et al. 2014; Bertsimas and Sim 2004), which decouples the uncertainties among different rows in the system matrix $\mathbf{H}$, i.e., denoting its i-th column as $\mathbf{h}_i$, with each column lying in an uncertainty set $\mathcal{S}_i$. We further relax the hard constraint $\mathbf{H}^T\boldsymbol{\mu} = \mathbf{0}$ to the following robust form:

$$\bar{\mathbf{h}}_i^T\boldsymbol{\mu}_i + \max_{\mathbf{h}_i\in\mathcal{S}_i} (\mathbf{h}_i - \bar{\mathbf{h}}_i)^T\boldsymbol{\mu}_i \leq \varepsilon_i, \tag{9}$$

where $\bar{\mathbf{h}}_i$ denotes the estimated nominal value of $\mathbf{h}_i$. For clarity, we illustrate our approach using the general polyhedral uncertainty set $\mathcal{S}_i = \{\boldsymbol{h}_i \mid \mathbf{D}_i\boldsymbol{h}_i \leq \boldsymbol{d}_i\}$, $i = 1, \cdots, N$, which has been widely used for robust optimization(Jiao, Yang, and Song 2025). Our method, however, readily extends to other common uncertainty sets such as ellipsoids and D-norms. Its generality is further demonstrated in our experiments on various uncertainty sets. As a result, the reformulated problem of (7) becomes:

$$\begin{aligned} v_t = & -\inf_{\mathcal{U}} \inf_{\boldsymbol{\mu}^{(t)},\mathbf{H}} [-\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{x})] \\ \text{s.t.} \quad & \begin{cases} \boldsymbol{\mu} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-, {\boldsymbol{\mu}^+}^T\boldsymbol{\mu}^- = \mathbf{0} \\ \rho_L\boldsymbol{u} \leq \boldsymbol{\mu}^+ + \boldsymbol{\mu}^- \leq \rho_U\boldsymbol{u}, \boldsymbol{u} \in \{0,1\}^N \\ u_m = 1 \text{ if } m \in \mathcal{U}, \text{ else } u_m = 0 \\ \bar{\mathbf{h}}_i^T\boldsymbol{\mu}_i + \max_{\mathbf{h}_i\in\mathcal{S}_i}(\mathbf{h}_i - \bar{\mathbf{h}}_i)^T\boldsymbol{\mu}_i \leq \varepsilon_i \\ \mathbf{D}_i\boldsymbol{h}_i \leq \boldsymbol{d}_i, \end{cases} \end{aligned} \tag{10}$$

where $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{x}) = \frac{1}{2\sigma^2}\left\{\|(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)\|_2^2 - 2(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)^T\boldsymbol{x}\right\}$

The problem remains challenging due to the presence of a maximization over an uncertainty set, which leads to a nested bilevel structure and the non-convex orthogonality constraint ${\boldsymbol{\mu}^+}^T\boldsymbol{\mu}^- = 0$. Thus, by denoting $\boldsymbol{p}_i$ the dual variable and leveraging strong duality (Yang et al. 2014), the maximization constraint under the polyhedron uncertainty set can be equivalently reformulated as:

$$\boldsymbol{p}_i^T\boldsymbol{d}_i \leq \varepsilon_i, \mathbf{D}_i^T\boldsymbol{p}_i = \boldsymbol{\mu}, \boldsymbol{p}_i \geq 0. \tag{11}$$

Introducing a binary auxiliary variable $\boldsymbol{b} \in \{0,1\}^N$ and using the upper bound ρ_U as a sufficiently large constant. The mutual exclusiveness between $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ at each index is replaced by the following two linear inequalities:

$$\boldsymbol{\mu}^+ + \rho_U \boldsymbol{b} \leq \rho_U \mathbf{1}, \boldsymbol{\mu}^- - \rho_U \boldsymbol{b} \leq \mathbf{0}. \tag{12}$$

Thus, (7) is reformulated as the following MIQP problem:

$$\begin{aligned} &\min \frac{1}{2\sigma^2} \left\{ \left\| (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-) \right\|_2^2 - 2(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)^T \boldsymbol{x} \right\} \\ &s.t. \quad \begin{cases} \boldsymbol{p}_i^T \boldsymbol{d}_i \leq \varepsilon_i, \mathbf{D}_i^T \boldsymbol{p}_i = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-, \forall i \\ \rho_L \boldsymbol{u} \leq \boldsymbol{\mu}^+ + \boldsymbol{\mu}^- \leq \rho_U \boldsymbol{u} \\ \boldsymbol{\mu}^+ + \rho_U \boldsymbol{b} \leq \rho_U \mathbf{1} \\ \boldsymbol{\mu}^- - \rho_U \boldsymbol{b} \leq \mathbf{0} \end{cases} \\ &\text{var} : \boldsymbol{u} \in \{0,1\}^M, \boldsymbol{b} \in \{0,1\}^M, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{p}_i \geq \mathbf{0}. \end{aligned} \tag{13}$$

Relaxation for Efficient Optimization

Mixed-integer problems are often solved using discrete methods like branch-and-bound, whose complexity grows rapidly with problem dimension. To improve efficiency, we reformulate the problem (13) via continuous relaxation. The tightness of the relaxation directly affects the approximation quality. While semidefinite programming (SDP) is a classical relaxation with tight bounds in combinatorial tasks such as max-cut (O'Donnell and Wu 2008) and graph coloring (Karger, Motwani, and Sudan 1998). Here, we propose a relaxation for problem (13),that can be theoretically proven to provide better bound than the traditional SDP relaxation.

Since the objective function of problem (13) is separable, here we first consider the one-dimensional case. By continuously relaxing b_m to $[0,1]$, we arrive at the following problem formulation.

$$\begin{aligned} &g(\mu_m^+, \mu_m^-, u_m, b_m, \{\boldsymbol{p}_i\}) = \\ &\begin{cases} 0, & \text{if } u_m = \mu_m^+ = \mu_m^- = 0, \\ f(\mu_m^+, \mu_m^-), & \text{if } u_m = 1, \rho_L \leq \mu_m^+ + \mu_m^- \leq \rho_H, \\ +\infty, & \text{otherwise}, \end{cases} \\ &(\mu_m^+, \mu_m^-, b_m, \{\boldsymbol{p}_i\}) \in \mathcal{P}, \end{aligned} \tag{14}$$

where f represents the objective function of (13), with feasible region $\mathcal{P}$ defined by all constraints excluding $\rho_L u_m \leq \mu_m^+ + \mu_m^- \leq \rho_H u_m$.

Then, we compute the convex envelope $\overline{co}(g)$ by constructing the convex hull in the epigraphical space of g (Frangioni and Gentile 2006, 2007), yielding:

$$\begin{aligned} &\overline{co}(g)(\mu_m^+, \mu_m^-, u_m, b_m, \{\boldsymbol{p}_i\}) \\ &= \begin{cases} 0, & \text{if } u_m = \mu_m^+ = \mu_m^- = 0 \\ h(\mu_m^+, \mu_m^-, u_m), & \text{if } u_m \in (0,1], \rho_L \leq \mu_m^+ + \mu_m^- \leq \rho_H, \\ +\infty, & \text{otherwise.} \end{cases} \\ &(\mu_m^+, \mu_m^-, b_m, \{\boldsymbol{p}_i\}) \in \mathcal{P} \end{aligned} \tag{15}$$

where $h(\cdot) = \frac{1}{2\sigma^2}\{u_m^{-1}(\mu_m^+ - \mu_m^-)^2 - 2x_m(\mu_m^+ - \mu_m^-)\}$ is a continuous relaxation of the objection in (13) and defining $0/0 := 0$. By introducing auxiliary variables ϕ_m such that $u_m^{-1}(\mu_m^+ - \mu_m^-)^2 \leq \phi_m$, which admits a second-order cone (SOC) formulation:

$$\begin{aligned} &\min \frac{1}{2\sigma^2} \sum_m \left\{ \phi_m - 2(\mu_m^+ - \mu_m^-)^T x_m \right\} \\ &s.t. \quad \begin{cases} \left\| \begin{matrix} \mu_m^+ - \mu_m^- \\ \frac{\phi_m - u_m}{2} \end{matrix} \right\| \leq \frac{\phi_m + u_m}{2}, \forall m \\ \boldsymbol{p}_i^T \boldsymbol{d}_i \leq \varepsilon_i, \mathbf{D}_i^T \boldsymbol{p}_i = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-, \forall i \\ \rho_L \boldsymbol{u} \leq \boldsymbol{\mu}^+ + \boldsymbol{\mu}^- \leq \rho_U \boldsymbol{u} \\ \boldsymbol{\mu}^+ + \rho_U \boldsymbol{b} \leq \rho_U \mathbf{1} \\ \boldsymbol{\mu}^- - \rho_U \boldsymbol{b} \leq \mathbf{0} \end{cases} \\ &\text{var} : \boldsymbol{u} \in [0,1]^N, \boldsymbol{b} \in [0,1]^N, \boldsymbol{\phi}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{p}_i \geq \mathbf{0}. \end{aligned} \tag{16}$$

Problem (16) is a Second-Order Cone Programming (SOCP) problem. We can demonstrate that this relaxation offers superior bound compared to traditional SDP relaxation.This is due to the absence of integer variables in the objective function, which leads to the degeneration of the SDP relaxation into a quadratic programming problem. Full derivations and equivalence proofs are deferred to the extended version (see the Links section).

Neural Unrolling for GPU-acceleration

Neural unrolling unfolds iterative optimization algorithms into trainable neural network layers, enabling efficient and adaptive solution approximation(Shi et al. 2021; He et al. 2022; Chen and Yang 2025). This approach naturally supports GPU acceleration, allowing parallel computation that significantly improves scalability and speed. We first construct the Lagrangian function of problem (16), as follows:

$$\begin{aligned} &\min \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{b}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{p}, \boldsymbol{\lambda}) \\ &= \min \frac{1}{2\sigma^2} \sum_m \left\{ \phi_m - 2(\mu_m^+ - \mu_m^-)^\top x_m \right\} \\ &\quad + \sum_\ell \left\{ \sum_j \lambda_j g_j(\cdot) \right\}_\ell, \\ &\quad \text{var} : \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{b}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{p}, \boldsymbol{\lambda}. \end{aligned} \tag{17}$$

where $g_j(\cdot)$ denotes the constraint functions and $\lambda_j > 0$ are the Lagrange multipliers, which acts as penalty coefficients.

The optimization of problem (17) proceeds by alternately updating the primal variables and the dual variables λ, which represent the penalty coefficients for constraint violations. By defining an RNN network, the optimization process can be represented by the parameter updates of the RNN network with K layers, as illustrated in Figure 1. The learnable parameters within each layer capture critical algorithmic components (e.g. step sizes). The optimization loss is defined based on objective of (17), guiding the training process. The iterative updates proceed until the difference between consecutive iterations falls below a threshold ϵ. Overall, the whole process of RoS-Guard is summarized in Algorithm 1.

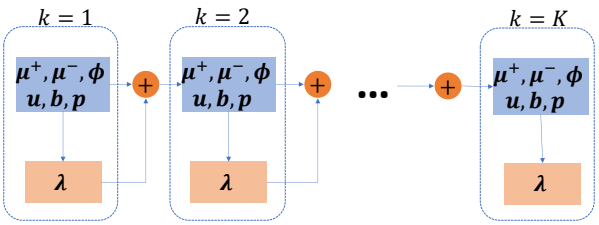

Figure 1: Illustration of the Unrolled Optimization Network

Algorithm 1: RoS-Guard algorithm

Input: $\{\boldsymbol{x}_t\}, \tau, \mathbf{H}, \rho_L, \rho_U, \sigma^2, \{\varepsilon_i\}, \{\mathcal{S}_i\}, h, \epsilon$
Initialize $t \leftarrow 0$, $V_0 \leftarrow 0$.
if *System is small-scale* **then**
  repeat
    $t \leftarrow t + 1$.
    Based on the relaxation (16), employ branch and bound method to solve problem (13), and thereby obtain the value of v_t.
    Update $V_t \leftarrow max\{V_{t-1}, 0\} + v_t$.
  until $V_t \geq h$;
  Output: $T_G \leftarrow t$, declare the change point T_G.
end
else if *System is large-scale* **then**
  repeat
    Initialize neural model parameters
    repeat
      optimize (17) with unrolling network
    until $|Loss_{prev} - Loss_{curr}| < \epsilon$;
    Obtain $v_t^{(\ell)}$ with optimized parameters and aggregate $v_t = \sum_{\ell=1}^{L} v_t^{(\ell)}$
    Update the decision statistic : V_t
  until $V_t \geq h$;
  Output: $T_G \leftarrow t$, declare the change point T_G.
end

Theoretical Analysis

As shown in Problem Statement (1) (2), OCD aims to minimize the worst-case ADD while satisfying a lower-bound constraint on the expected false alarm period (FAP). In this section, we provide a theoretical analysis of our proposed algorithm by establishing performance guarantees. Specifically, we first derive a sufficient condition to ensure the expected FAP constraint is met, and then present an upper bound on the worst-case ADD under any given detection threshold.

Condition to Meet Expected False Alarm Period Constraint

According to (5), the change is declared at the first time $V_K \geq h$. Therefore, the value of the threshold h holds great significance in the detection process. In this subsection, we derive a sufficient condition for our method. For any prescribed lower bound γ on the FAP, this condition provides a guideline for the selection of h to ensure that the FAP constraint can be satisfied.

Theorem 1. *Suppose, in general(Jiao, Yang, and Jian 2025), that $\mathbf{x}^{(t)}$ is upper bounded, i.e, $\|\mathbf{x}^{(t)}\|_2^2 \leq \alpha$. The expected false alarm period of RoS-Guard is greater than γ if:*

$$h \geq \frac{\alpha}{2\sigma^2}\gamma. \tag{18}$$

Proof Sketch. Recall that Γ_R represents the stopping time of RoS-Guard as shown in (5). The core of proof resides in establishing the relationship between $\mathbb{E}_\infty\{V_{\Gamma_R}\}$ and $\mathbb{E}_\infty\{\Gamma_R\}$, which is achieved by the upper bound on v_t. When the false alarm is declared, we have $\mathbb{E}_\infty\{V_{\Gamma_R}\} \geq h$, which thereby elucidates the relationship between the $\mathbb{E}_\infty\{\Gamma_R\}$ and h. The complete proof can be found in in the extended version. □

Upper Bound on the Worst-Case Expected Detection Delay

For OCD task, it is imperative to evaluate the worst-case expected detection delay of the detector. In this subsection, we provide the theoretical analysis of the worst-case expected detection delay of the proposed method. Define $J(\Gamma_R)$ as Lorden's worst-case expected detection delay. The subsequent theorem concerning the worst-case expected detection delay can be obtained.

Theorem 2. *Let d_i represent the diameter of uncertainty set $\mathcal{U}_i$. For any threhold h, by employing Wald's approximations (Tartakovsky, Nikiforov, and Basseville 2014b), set $\varepsilon_i \geq d_i \rho_H M^{\frac{1}{2}}$, the worst-case expected detection delay of Ros-Gurad can be bounded as follows,*

$$J(\Gamma_R) \leq \frac{2h\sigma^2}{{\rho_L}^2}, \tag{19}$$

Proof Sketch. We begin by introducing a lower bound on the expectation of v_t when change occurs at a time instant t_a based on the bounds of $\boldsymbol{\mu}$. Subsequently, we proceed to derive an upper bound on $\mathbb{E}_{t_a}\{(\Gamma_R - t_a + 1)^+ \mid \mathcal{F}_{t_a-1}\}$ for the proposed method. Next, we demonstrate that the proposed method achieves the equalizer rule given the pre- and post-change model elaborated in (3), Finally, leveraging these insights and employing Wald's approximations, we substantiate Theorem 2. The complete proof can be found in the extended version (see the Links section). □

Experiments

To evaluate the effectiveness of our proposed RoS-Guard algorithm, we conduct experiments on two representative online change detection scenarios: (i) attack injection detection in smart grids, and (ii) channel blockage detection in MIMO wireless systems. We compare our method against two state-of-the-art baselines, RGCUSUM and CyberQCD, which are compatible with our system modeling. Furthermore, to assess the scalability of RoS-Guard in different scale systems, we measure and compare the detection latency under varying observation dimensions.

Datasets

Dataset I: We follow (Huang et al. 2011) to formulate FDIA detection in smart grids using the dynamic DC power flow model in (3). The system includes $N+1$ buses and M meters, with state $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^N$ and observations $\mathbf{x}^{(t)} \in \mathbb{R}^M$. The measurement matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$ is determined by grid topology and line susceptance (Kosut et al. 2011; Cui et al. 2012). We conduct experiments using the IEEE-14 bus system, a standard benchmark. The system state is initialized using "case14" in MATPOWER (Zimmerman and Murillo-Sánchez 2016), and we simulate attacks by injecting random vectors into $\mathbf{x}^{(t)}$ from time t_a. Following (Li, Yılmaz, and Wang 2014), we assume the measurement matrix is fully known except in one sub-region. Uncertainty sets and perturbed $\mathbf{H}$ are detailed in the extended version (see the Links section).

Dataset II: MIMO technology leverages multiple antennas at both the transmitter and receiver to enhance transmission efficiency via multiple signal paths (Tsoulos 2018). However, these paths are susceptible to blockages from static or dynamic obstacles such as buildings, vehicles, or humans (Vaigandla and Nookala 2021). Let $\mathbf{x}^{(t)} \in \mathbb{R}^M$ be the received signal vector and $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^N$ the transmitted signal vector, where M and N are the numbers of receive and transmit antennas, respectively. The channel matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$ contains the channel gains between antenna pairs. Due to environmental uncertainties, these coefficients are typically imprecise. We adopt a 2×4 MIMO testbed based on USRP devices. As shown in the extended version (see the Links section), various blockage scenarios are considered. For each case, we continuously collect the received signals and apply OCD detectors to identify blockage events.

Performance Evaluation

In the introduction, we underscored the limitations of conventional OCD methods due to their reliance on certain assumptions. These assumptions encompass constant distributions before and after a change, as well as specific conditions for the system state. Consequently, these methods prove unsuitable for addressing the problem under investigation. We implemented the Adaptive CUSUM algorithm (Huang et al. 2011) In our experiments to exemplify this fact. Furthermore, a performance comparison is conducted between our algorithm, the method proposed in (Li, Yılmaz, and Wang 2014) (referred to as CyberQCD for simplicity), and the RGCUSUM algorithm (Zhang and Wang 2021), which claims to be the state-of-the-art detector for the OCD in dynamic systems. To better showcase the effectiveness and superiority of the proposed method, we approximating the value of v_t by directly using the optimal objective function value of (16).

After executing each detector 500 runs, the experimental results on data I and data II are visually presented in Figure 2 and Figure 3.

Following conventional metrics for online change detection, we evaluate the ADD of different detectors under the same FAP. Recall that FAP signifies the point at which the detector stops when no change is detected, thus serving as a measure to assess the risk of false alarms. It is evident from Figure 2 (a) and Figure 3 that our method consistently exhibits a smaller average detection delay for any given FAP, which emphasizes the superior performance achieved by RoS-Guard.

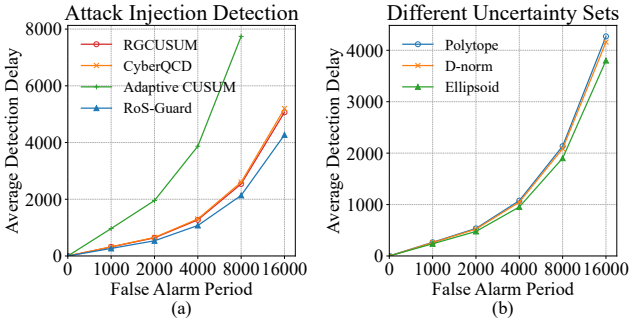

Figure 2: (a) Performance Comparison with Attack Injection Detection in Smart Grids System, (b) Performance Comparison with Various Uncertainty Sets.

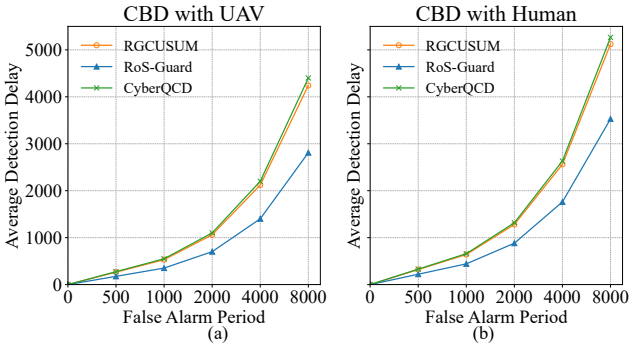

Figure 3: Performance Comparison with Blockage Detection(CBD) in Wireless MIMO System.

From Figure 2(a), we can observe that the performance of the adaptive CUSUM algorithm is unacceptable. This can primarily be attributed to the inconsistency between its assumptions and the experiment setup. Specifically, the adaptive CUSUM algorithm assumes a Gaussian distribution for the systems state, which does not hold in our experiments. RGCUSUM and CyberQCD show comparable performance, with both falling short in comparison to the performance of the RoS-Guard. This result is in accordance with our expectation, as both the RGUCUSM and CyberQCD methods are developed based on the perfectly known $\mathbf{H}$. When the system matrix $\mathbf{H}$ is inaccurate, their detection performance rapidly deteriorates. In contrast, the RoS-Guard algorithm effectively handles uncertainties within the system matrix, ensuring robust and reliable detection performance. Furthermore, we conducted additional experiments to showcase the performance of the RoS-Guard algorithm when the system matrix $\mathbf{H}$ is assumed to belong to different uncertainty sets. These uncertainty sets, which are commonly employed in practical applications, include ellip-

soid uncertainty sets, D-norm uncertainty sets, and polyhedron uncertainty sets (Yang et al. 2014). Detailed information regarding the settings of these uncertainty sets can be found in the extended version (see the Links section). The experiments were conducted with a total of 500 Monte Carlo runs. As depicted in Figure 2 (b), the RoS-Guard algorithm consistently demonstrates superior performance across diverse uncertainty sets, underscoring its remarkable ability to generalize and adapt to varying conditions.

Evaluation of GPU Parallel Computing

To compare the detection time differences between CPU- and GPU-based algorithms under varying system scales, we follow the setup of Dataset I and consider attack injection under system observations of size 2^k . By tuning parameters to ensure comparable detection performance, we evaluate and compare the execution times of the CPU and GPU algorithms. The average runtime is computed over 50 randomized trials. The detailed experimental settings are provided in the extended version (see the Links section).

Experimental result in Figure 4 shows that our GPU-based neural unrolling algorithm exhibits a significant speed advantage when the system scale is large. Specifically, when the system reaches a scale of 2^8, the GPU-based parallel algorithm achieves more than a 20× speedup.

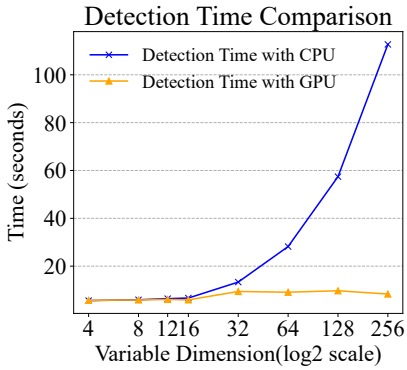

Figure 4: Detection Time Comparison between Neural Unrolling with GPU and MIQP with CPU in Different System Scale.

It is worth noting that if we slightly violate the online change detection setting by grouping a small number T of consecutive observations into a batch, the GPU-based method does not incur additional execution time. This indicates that when the algorithm's execution time is comparable to the system's observation interval, the efficiency advantage of the GPU-based method scales approximately by a factor of T

Numerical Verification of Theorem 1 and Theorem 2

In this subsection, we numerically verify Theorem 1 and Theorem 2 on Data I. The number of Monte Carlo runs is 100. In order to verify Theorem 1, for any prescribed γ, we set the value of h to be the lower bound calculated from the right side of inequality in (18). And then we evaluate the actual average FAP of RoS-Guard with the same h. The results are shown in Figure 5(a). It is seen that under every setting of (ρ_L, ρ_H), the actual FAP is always larger than γ for our detector, which confirms Theorem 1.

To demonstrate the validity of Theorem 2, we compute the upper bounds on $J(\Gamma_R)$ for different h values by employing (19). And then for each value of h, we numerically calculate the corresponding actual ADD of the proposed method. The experimental results are shown in Figure 5(b). It is seen that the actual ADD of RoS-Guard is consistently smaller than the upper bounds on $J(\Gamma_R)$, which validates Theorem 2.

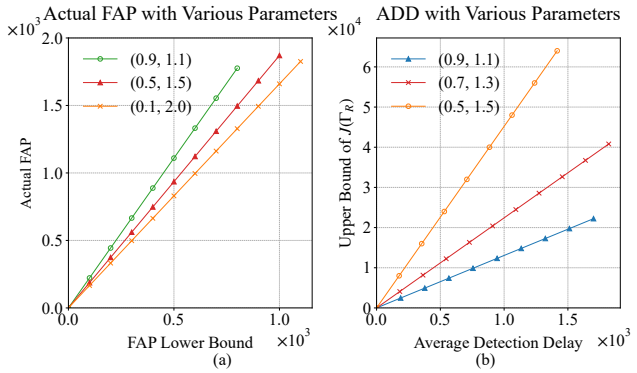

Figure 5: Numerical Verification of Theorems

Conclusion

In real-world applications, system uncertainty caused by estimation errors, model drift, and environmental disturbances presents a fundamental challenge to online change detection. To address this challenge, we propose RoS-Guard, a robust and scalable detection framework. Our method explicitly addresses system uncertainty. Based on Lorden's minimax delay formulation, it achieves the minimax-optimal worst-case detection delay when solved exactly, while providing near-optimal performance in practical implementations. To enable deployment in large-scale systems, we develop a parallel algorithm leveraging neural unrolling with GPU acceleration, which significantly improves computational efficiency without sacrificing accuracy. We provide theoretical guarantees on detection performance, and extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness, scalability, and robustness of the proposed approach.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 12371519 and 61771013; in part by Asiainfo Technologies; in part by the Fundamental Research Funds for the Central Universities of China; and in part by the Fundamental Research Funds of Shanghai Jiading District.

References

Bertsimas, D.; and Sim, M. 2004. The price of robustness. *Operations research*, 52(1): 35–53.

Chen, X.; and Yang, K. 2025. GPU-accelerated parallel bilevel optimization for roubst 6g isac. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11167–11175.

Cui, S.; Han, Z.; Kar, S.; Kim, T. T.; Poor, H. V.; and Tajer, A. 2012. Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions. *IEEE Signal Processing Magazine*, 29(5): 106–115.

Dou, S.; Yang, K.; and Poor, H. V. 2019. PC2A: Predicting Collective Contextual Anomalies via LSTM With Deep Generative Model. *IEEE Internet of Things Journal*, 6(6): 9645–9655.

Frangioni, A.; and Gentile, C. 2006. Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106: 225–236.

Frangioni, A.; and Gentile, C. 2007. SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Operations Research Letters*, 35(2): 181–185.

Ghavami, K.; and Naraghi-Pour, M. 2017. MIMO detection with imperfect channel state information using expectation propagation. *IEEE Transactions on Vehicular Technology*, 66(9): 8129–8138.

Gopalan, A.; Lakshminarayanan, B.; and Saligrama, V. 2021. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems*, 34: 29064–29073.

Hare, J. Z.; and Kaplan, L. 2022. Uncertainty-Aware Quickest Change Detection: An Experimental Study. In *2022 25th International Conference on Information Fusion (FUSION)*, 1–8. IEEE.

Hare, J. Z.; Kaplan, L.; and Veeravalli, V. V. 2021. Toward uncertainty aware quickest change detection. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, 1–8. IEEE.

He, S.; Xiong, S.; An, Z.; Zhang, W.; Huang, Y.; and Zhang, Y. 2022. An unsupervised deep unrolling framework for constrained optimization problems in wireless networks. *IEEE Transactions on Wireless Communications*, 21(10): 8552–8564.

Huang, Y.; Li, H.; Campbell, K. A.; and Han, Z. 2011. Defending false data injection attack on smart grid network using adaptive CUSUM test. In *2011 45th Annual Conference on Information Sciences and Systems*, 1–6. IEEE.

Huang, Y.; Tang, J.; Cheng, Y.; Li, H.; Campbell, K. A.; and Han, Z. 2014. Real-time detection of false data injection in smart grid networks: An adaptive CUSUM method and analysis. *IEEE Systems Journal*, 10(2): 532–543.

Jiao, Y.; Yang, K.; and Jian, C. 2025. DTZO: Distributed Trilevel Zeroth Order Learning with Provable Non-Asymptotic Convergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 27873–27916. PMLR.

Jiao, Y.; Yang, K.; and Song, D. 2025. Federated distributionally robust optimization with non-convex objectives: Algorithm and analysis. *IEEE Transactions on Mobile Computing*.

Karger, D.; Motwani, R.; and Sudan, M. 1998. Approximate graph coloring by semidefinite programming. *Journal of the ACM (JACM)*, 45(2): 246–265.

Kosut, O.; Jia, L.; Thomas, R. J.; and Tong, L. 2011. Malicious data attacks on the smart grid. *IEEE Transactions on Smart Grid*, 2(4): 645–658.

Li, K.; Cao, X.; and Meng, D. 2024. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Li, S.; and Wang, X. 2015. Cooperative change detection for voltage quality monitoring in smart grids. *IEEE Transactions on Information Forensics and Security*, 11(1): 86–99.

Li, S.; Yılmaz, Y.; and Wang, X. 2014. Quickest detection of false data injection attack in wide-area smart grids. *IEEE Transactions on Smart Grid*, 6(6): 2725–2735.

Lorden, G. 1971. Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, 1897–1908.

Molloy, T. L.; and Ford, J. J. 2017. Misspecified and asymptotically minimax robust quickest change detection. *IEEE Transactions on Signal Processing*, 65(21): 5730–5742.

O'Donnell, R.; and Wu, Y. 2008. An optimal SDP algorithm for Max-Cut, and equally optimal Long Code tests. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 335–344.

Pei, Y.; Liang, Y.-C.; Teh, K. C.; and Li, K. H. 2011. Secure communication in multiantenna cognitive radio networks with imperfect channel state information. *IEEE Transactions on Signal Processing*, 59(4): 1683–1693.

Shi, Y.; Choi, H.; Shi, Y.; and Zhou, Y. 2021. Algorithm unrolling for massive access via deep neural networks with theoretical guarantee. *IEEE Transactions on Wireless Communications*, 21(2): 945–959.

Tartakovsky, A.; Nikiforov, I.; and Basseville, M. 2014a. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press.

Tartakovsky, A.; Nikiforov, I.; and Basseville, M. 2014b. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press.

Tsoulos, G. 2018. *MIMO system technology for wireless communications*. CRC press.

Unnikrishnan, J.; Veeravalli, V. V.; and Meyn, S. P. 2011. Minimax robust quickest change detection. *IEEE Transactions on Information Theory*, 57(3): 1604–1614.

Vaigandla, K. K.; and Nookala, V. 2021. Survey on Massive MIMO: Technology, Challenges, Opportunities and Benefits. Ssrn working paper, YMER.

Weber, T.; Sklavos, A.; and Meurer, M. 2006. Imperfect channel-state information in MIMO transmission. *IEEE Transactions on Communications*, 54(3): 543–552.

Xie, L. 2022. Minimax robust quickest change detection using wasserstein ambiguity sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*, 1909–1914. IEEE.

Yang, K.; Huang, J.; Wu, Y.; Wang, X.; and Chiang, M. 2014. Distributed robust optimization (DRO), part I: Framework and example. *Optimization and Engineering*, 15(1): 35–67.

Yang, K.; Liu, R.; Sun, Y.; Yang, J.; and Chen, X. 2016. Deep network analyzer (DNA): A big data analytics platform for cellular networks. *IEEE Internet of Things Journal*, 4(6): 2019–2027.

Yu, W.; Zhang, X.; Das, S.; Zhu, X. X.; and Ghamisi, P. 2024. Maskcd: A remote sensing change detection network based on mask classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Zhang, J.; and Wang, X. 2021. Low-complexity quickest change detection in linear systems with unknown time-varying pre-and post-change distributions. *IEEE Transactions on Information Theory*, 67(3): 1804–1824.

Zhang, Q.; Sun, Z.; Herrera, L. C.; and Zou, S. 2024. Data-Driven Quickest Change Detection in (Hidden) Markov Models. *IEEE Transactions on Signal Processing*.

Zimmerman, R. D.; and Murillo-Sánchez, C. E. 2016. Matpower 6.0 user's manual. *Power Systems Engineering Research Center*, 9.