Robust Estimation in Fully Distributed Sensor Network in the Presence of Outliers

Hardik Aggarwal

Master of Science Thesis

TU Delft Delft University of Technology

Delft Center for Systems and Control

Robust Estimation in Fully Distributed Sensor Network in the Presence of Outliers

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Hardik Aggarwal

July 24, 2025

Faculty of Mechanical Engineering (ME) · Delft University of Technology

Abstract

Classical distributed estimation algorithms for state estimation in Wireless Sensor Networks (WSNs), including consensus-based Kalman filtering and diffusion strategies, typically assume Gaussian observation models under which outliers are rare. However, even a small fraction of outliers can significantly corrupt local updates and propagate errors through the network, degrading the global estimation performance. Furthermore, the distributed setting imposes constraints such as unreliable measurements and limited communication resources, necessitating robust and computationally efficient estimation techniques. This paper proposes a fully distributed estimation framework that integrates a convex, smooth log-cosh loss function within a generalized Bayesian inference formulation to enable outlier-resilient state estimation. The resulting robust update is embedded into a recursive filtering structure and solved via the Exact First Order Algorithm (EXTRA) algorithm for distributed consensus optimization, eliminating the need for dual variables or inner-loop minimization. A formal stability analysis is conducted by conservatively modeling the estimator as a Kalman Filter (KF) with intermittent observations, and a sufficient condition on the robustness parameter is derived to guarantee bounded mean-square error. To improve performance without violating this stability condition, an adaptive strategy is introduced to dynamically adjust the robustness parameter based on the residual magnitude. Theoretical analysis and numerical simulations demonstrate that the proposed method achieves accurate and resilient estimation in the presence of impulsive noise and adversarial disturbances, reducing average Root Mean Square Error (RMSE) by 5-10% under measurement outliers and up to 20% in scenarios with process disturbances, while also requiring fewer consensus iterations for convergence.

Table of Contents

Preface

This thesis is the final deliverable of my Master's degree in Systems and Control. It addresses the significant issue of robust state estimation within fully distributed sensor networks, particularly when confronted with measurement outliers that can corrupt the estimation process.

The primary technical derivations and core contributions of this research are comprehensively presented in the self-contained paper found in Chapter 4. This paper details the proposed log-cosh-based estimation framework, its implementation via the EXTRA consensus algorithm, the formal stability analysis, and empirical validation.

Therefore, this thesis is best viewed as an introduction and companion to the research paper. The preceding chapters provide the broader context, motivation, and foundational background on distributed estimation and consensus algorithms that could not be included in the concise format of the paper. A significant aspect of this thesis is the detailed comparative analysis of various robust loss functions, which serves to justify the specific design choices made in the paper. While there is some intentional overlap, the thesis provides the foundational framework that motivates and introduces the main work.

Acknowledgements

I am grateful to my supervisor, Dr. Nitin Myers, for giving me the chance to work on this thesis at the Delft Center for Systems and Control.

I would like to sincerely thank my daily supervisor, Dr. Chen Quan. The weekly meetings were instrumental in keeping me focused. Her guidance and feedback were crucial in driving my progress and were essential to the success of my thesis.

To my friends at TU Delft and DCSC: I am grateful to each of you for the enjoyable moments, the late-night study sessions, and for the chance to learn so much in your company.

To my family: Mom, Dad, and my Sister. Your support has made all of this possible, and I am endlessly grateful to you.

Delft, University of Technology
July 24, 2025

Hardik Aggarwal

Chapter 1

Introduction

1-1 Motivation

Wireless Sensor Networks (WSNs) play a crucial role in applications like environmental monitoring, structural health diagnostics, and autonomous surveillance, where estimating the state from measurements spread across different locations is crucial for making timely decisions and exercising control. [1, 2]. Traditional centralized architectures aggregate sensor data at a Fusion Center (FC) and employ optimal estimators such as the Kalman Filter (KF) [3]. However, these centralized designs are severely constrained by communication bottlenecks, single-point failures, and scalability limitations, particularly in dynamic or large-scale deployments.

To mitigate these limitations, distributed estimation algorithms based on consensus [4, 5], diffusion [6], and gossip [7] strategies have been developed. These frameworks enable sensor nodes to perform local updates and refine their state estimates through direct communication with neighbors. Distributed methods are inherently scalable and robust to link or node failures, making them suitable for practical WSNs operating under energy, bandwidth, and topology constraints.

However, a critical vulnerability remains: robustness against outliers [8]. Under real-world conditions, sensor readings are often corrupted by faults, adversarial attacks, or impulsive noise with heavy-tailed distributions. These anomalous measurements can propagate through iterative update schemes, contaminating the global state estimates and compromising reliability [9]. Standard KF-based approaches assume Gaussian noise and disproportionately penalize large deviations, making them ill-equipped for such conditions. Even recent distributed variants, such as the Distributed Kalman Filter (DKF) [5, 10], Generalized Kalman Consensus Filter (GKCF) [11], or consensus filters [12], implicitly rely on light-tailed noise models and fail to suppress the influence of extreme observations.

Robust estimation techniques have been proposed to address these problems. For instance, the Robust Consensus Nonlinear Information Filter (RCNIF) proposed in [13] employs a

Student-t model within a variational Bayesian framework, and the Diffusion Minimum Generalized Rank Norm (dMGRN) proposed in [14] adopts rank-norm statistics with Minimum Volume Ellipsoid (MVE)-based weighting. Although effective, these methods are computationally intensive and incompatible with resource-constrained nodes. Alternatively, the centralized Weighted Observation Likelihood Filter (WoLF) algorithm proposed in [15] leverages generalized Bayesian inference using reweighted log-likelihood to achieve robustness with closed-form updates. However, its centralized nature renders it unsuitable for distributed applications. This motivates the design of a robust distributed estimation algorithm that avoids these limitations. The key challenge is enabling each sensor node in a WSN to estimate global system state with outlier-contaminated measurements using local computations and neighbor communication while ensuring stability and accuracy.

1-2 Objectives

1-2-1 Problem Statement

Despite significant advances in distributed estimation, a fundamental gap remains in the design of algorithms that are robust to heavy-tailed noise while being scalable, stable, and implementable on resource-constrained sensor nodes. Specifically, most existing methods either

- assume light-tailed noise models, making them vulnerable to extreme outliers;
- rely on centralized architectures or fusion centers, limiting scalability and fault tolerance;
- or require computationally demanding optimization techniques that are infeasible for embedded or battery-powered platforms.

This thesis aims to design a robust and fully distributed estimation algorithm that addresses these limitations. Specifically, the following question is addressed:

How can each sensor node in a fully distributed WSN estimate the global system state in the presence of outlier-contaminated measurements using only local computations and communication with neighbors, while guaranteeing stability and accuracy of the resulting estimates?

1-2-2 Research Questions

- How can a robust and computationally efficient state estimation algorithm be designed for fully distributed WSNs to effectively handle outliers and heavy-tailed noise?
- Can a log-cosh loss function be integrated within a generalized Bayesian framework to implement a recursive filter that not only provides resilience to outliers but also comes with formal stability guarantees?

1-2-3 Technical Contribution

This thesis introduces a fully distributed robust filtering framework that integrates generalized Bayesian inference with consensus optimization. The core innovations are:

- **Robust loss formulation:** A smooth and convex log-cosh penalty replaces the standard quadratic loss on whitened residuals, enabling robustness to heavy-tailed errors and still maintaining KF like optimality for when there are no outliers.
- **EXTRA-based consensus optimization:** The proposed method leverages the Exact First Order Algorithm (EXTRA) algorithm proposed in [16] for distributed minimization of the log-cosh-penalized objective, allowing nodes to converge to a consistent Maximum A Posteriori (MAP) estimate using only local communication and gradient updates.
- **Theoretical stability guarantee:** A theoretical stability guarantee is derived, along with a conservative upper bound on the robustness parameter in the loss function, which governs the trade-off between sensitivity to outliers and estimation accuracy, thereby ensuring boundedness of the estimation error covariance.
- **Dynamic robustness adjustment:** An adaptive mechanism is introduced to tune the robustness parameter in real time based on local residuals, improving filter performance without compromising robustness under outliers.

Collectively, these contributions form a robust, stable, and scalable distributed estimation architecture that is well suited for deployment in real-world WSNs applications with adversarial or faulty sensing environments.

1-3 Outline

The remainder of this thesis is organized as follows.

Chapter 2 offers the foundational information needed to understand this thesis. It includes the discussion of distributed estimation WSNs, graph representation of the network, consensus algorithms, which are crucial for distributed estimation, and the centralized Kalman filter is examined as a baseline.

Chapter 3 focuses on robust estimation techniques designed to handle outliers. It reviews existing frameworks, such as Student-t-based filters and covariance inflation, before introducing the proposed filter, which is based on a log-cosh loss function. This chapter provides a detailed comparison with several alternative loss functions to justify the selection of log-cosh.

Chapter 4 presents the core technical contributions of this thesis in the form of a self-contained research paper. The paper details the proposed robust estimation framework, its implementation in a distributed network using the EXTRA algorithm, a formal stability analysis, and comprehensive simulation results.

Chapter 5 concludes and summarizes this thesis. It also discusses some limitations and suggests directions for future research.

Chapter 2

Background

2-1 Distributed Estimation in Sensor Networks

The goal of distributed state estimation is to enable all spatially distributed nodes to collaboratively compute an accurate estimate of the global system state over time, using only local observations and information exchanged with neighboring nodes.

Consider a network of N sensor nodes, indexed by the set $\mathcal{V} = \{1, \ldots, N\}$, observing a discrete-time linear dynamical system described by

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{w}_k, \tag{2-1}$$

$$\mathbf{z}_{i,k} = \mathbf{H}_{i,k} \mathbf{x}_k + \boldsymbol{\nu}_{i,k}, \quad i \in \mathcal{V}, \tag{2-2}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ denotes the global state vector at time k, and $\mathbf{z}_{i,k} \in \mathbb{R}^{m_i}$ is the measurement acquired by sensor node i. The matrices $\mathbf{A}_k \in \mathbb{R}^{n \times n}$ and $\mathbf{H}_{i,k} \in \mathbb{R}^{m_i \times n}$ define the state transition and local observation models, respectively.

The process noise $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ is assumed to be zero-mean Gaussian with covariance $\mathbf{Q}_k \succ \mathbf{0}$, independent across time. In contrast, the measurement noise $\boldsymbol{\nu}_{i,k}$ may follow an arbitrary distribution to account for non-Gaussian or heavy-tailed behavior, such as impulsive noise or sensor faults. The measurement noise is assumed independent across time and sensors (i.e., uncorrelated between different nodes $i \neq j$) while also uncorrelated with the process noise. This thesis considers measurement noise to be heavy-tailed, unimodal, and symmetric, which can be approximated by a Gaussian with covariance $\mathbf{R}_{i,k} \succ \mathbf{0}$ for algorithmic tractability.

Distributed estimation addresses the limitations of centralized data collection by a Fusion Center (FC) by enabling each node to maintain its own local estimate of the overall state, which is then progressively refined through interactions with neighboring nodes. Common architectures alternate between two phases:

- **Local prediction and update:** Each node first propagates its prior through the process model (prediction), then assimilates its own measurement (update) using a Bayesian filter, e.g., a local Kalman Filter (KF).

- **Consensus or fusion:** Nodes exchange intermediate data, state means, information vectors, or covariances, with their neighbors and fuse this information to approximate the global posterior.

This fully distributed method provides scalability, fault tolerance to single-node failures, and reduced communication costs relative to centralized alternatives. The corresponding distributed estimation strategies have been extensively explored under diverse conditions, including static or time-varying network topologies, linear or nonlinear dynamics, and Gaussian or non-Gaussian noise.

2-2 Graph Representation of Sensor Networks

Graph theory provides a natural mathematical framework for modeling the underlying network topology, characterizing how sensors are interconnected and exchange information, and for analyzing its impact on estimation quality, convergence speed, and robustness [4].

2-2-1 Network Model

We model a wireless sensor network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where:

- $\mathcal{V} = \{1, 2, \ldots, N\}$ is the set of sensor nodes.
- $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of bidirectional communication links.

An edge $(i, j) \in \mathcal{E}$ exists if nodes i and j can reliably exchange messages within a single time step. This capability depends on physical factors such as transmission range, interference, and obstacles in the environment. We assume that the graph stays connected during the process, enabling information from any node to be transmitted to all other nodes through multi-hop communication. The closed neighborhood of node i, including itself, is defined as

$$\mathcal{N}_i = \{j \in \mathcal{V} : (i, j) \in \mathcal{E}\} \cup \{i\}. \tag{2-3}$$

Each node collects local measurements and communicates only with its immediate neighbors to collaboratively estimate the system state. The structure and connectivity of the network influence key performance aspects, such as information propagation speed and algorithm robustness, which we explore through associated matrices and graph properties.

2-2-2 Matrix Representations

Matrix representations of the underlying network topology are discussed next to support algorithmic design and convergence analysis.

Adjacency Matrix

The unweighted adjacency matrix $\mathbf{A}_{\text{adj}} \in \mathbb{R}^{N\times N}$ encodes the communication topology as

$$[\mathbf{A}_{\text{adj}}]_{ij} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{2-4}$$

Since the network contains no self-loops, all diagonal entries satisfy $[\mathbf{A}_{\text{adj}}]_{ii} = 0$.

For weighted graphs, non-negative weights w_{ij} are assigned to edges to reflect factors such as communication reliability or signal strength. These weights define a weighted adjacency matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{N\times N}$, satisfying

$$w_{ij} \geq 0, \quad w_{ij} = 0 \text{ if } (i,j) \notin \mathcal{E} \text{ and } i \neq j, \quad w_{ii} > 0. \tag{2-5}$$

The positive diagonal entries w_{ii} allow the nodes to incorporate their own information during updates. In consensus-based protocols, $\mathbf{W}$ is often designed to be row-stochastic (rows sum to 1) or doubly stochastic (both rows and columns sum to 1).

$$\sum_{j=1}^{N} w_{ij} = \sum_{i=1}^{N} w_{ij} = 1, \quad \forall i \in \mathcal{V}. \tag{2-6}$$

This ensures balanced information fusion and convergence to a global average.

Common weight assignment strategies include the following:

- **Uniform:** $w_{ij} = 1/|\mathcal{N}_i|$ for all $j \in \mathcal{N}_i$, including i (resulting in $w_{ii} = 1/|\mathcal{N}_i|$).
- **Metropolis-Hastings:**

$$w_{ij} = \begin{cases} \frac{1}{1+\max\{d_i,d_j\}}, & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j, \\ 1 - \sum_{j\in\mathcal{N}_i\setminus\{i\}} w_{ij}, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{2-7}$$

  where $d_i = |\mathcal{N}_i| - 1$ is the degree of node i. This scheme ensures symmetry and local computability and prevents high-degree nodes from dominating, thereby enhancing robustness in heterogeneous networks [17].

- **Distance-based or reliability-based:** $w_{ij} \propto 1/d_{ij}^{\alpha}$ or based on signal-to-noise ratio for $j \in \mathcal{N}_i \setminus \{i\}$, normalized to satisfy stochasticity conditions. This assigns higher weights to the more reliable links.

The choice of $\mathbf{W}$ directly affects the convergence speed and stability in distributed estimation.

Degree Matrix and Graph Laplacian

The degree matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose entries represent the number of neighbors of each node:

$$[\mathbf{D}]_{ii} = d_i = |\mathcal{N}_i| - 1 = \sum_{j=1}^{N} [\mathbf{A}_{\text{adj}}]_{ij}. \tag{2-8}$$

The graph Laplacian $\mathbf{L} \in \mathbb{R}^{N \times N}$ is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}_{\text{adj}}, \tag{2-9}$$

where $\mathbf{A}_{\text{adj}}$ is the unweighted adjacency matrix. The Laplacian is symmetric and positive semidefinite. It plays a central role in analyzing convergence properties of distributed algorithms, such as consensus protocols.

The spectrum of $\mathbf{L}$, denoted by its real eigenvalues $0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L})$, provides key insights. The smallest eigenvalue is always $\lambda_1(\mathbf{L}) = 0$, with associated eigenvector $\mathbf{1} \in \mathbb{R}^N$ (the all-ones vector). The second-smallest eigenvalue, $\lambda_2(\mathbf{L}) > 0$, is known as the *algebraic connectivity* [18]. It indicates graph connectivity: the graph is connected if and only if $\lambda_2(\mathbf{L}) > 0$. Larger values of $\lambda_2(\mathbf{L})$ imply stronger connectivity and faster convergence rates (discussed in the next section) because they facilitate quicker information diffusion.

2-3 Consensus Algorithms

Consensus algorithms form the mathematical foundation of distributed estimation, enabling nodes to collaboratively compute global quantities through local interactions in the fully distributed Wireless Sensor Networks (WSNs) [4]. In distributed Kalman filtering, consensus protocols allow nodes to fuse local state estimates without centralized coordination. This approach is useful in fully distributed WSNs, where nodes operate autonomously while achieving collective goals, such as estimating environmental parameters or tracking moving targets, offering both scalability and resilience to unpredictable changes in network topology.

2-3-1 Motivation and Basic Principle

Consider a simple problem with N sensor nodes in the network, each initially holding an observation value, a scalar $x_i^{(0)}$, and the collective objective is for all nodes to compute the network-wide average

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i^{(0)}$$

using only neighbor-to-neighbor communication, without any central authority collecting or processing the data. Although seemingly simple, this task captures the essence of distributed information fusion because it requires nodes to share and aggregate local data to mimic centralized computation. Consensus algorithms are specifically designed to address this challenge. Consensus algorithms are inspired by the principle of emergent coherence seen in many natural systems, such as the synchronized flashing of fireflies or the flocking patterns of birds, where global order arises from simple, local rules. In the context of sensor networks, this principle is translated into a simple iterative update rule: each sensor adjusts its current estimate based on the information from its local neighborhood. This repeated act of local averaging effectively diffuses information throughout the graph. As a result, even without any global oversight, the entire network can converge to a single, consistent state. Moreover, this averaging problem can be extended to complex scenarios in distributed filtering, such as combining local state estimates to form a global view, fusing covariance matrices to quantify uncertainty, or averaging quantities such as sensor measurements or gradient information in optimization-based estimators.

2-3-2 Mathematical Formulation

Discrete-Time Average Consensus

In distributed sensor networks with synchronized communication cycles, consensus algorithms operate in discrete time. At each iteration ℓ, node i updates its local state using the weighted averaging rule:

$$x_i^{(\ell+1)} = w_{ii}x_i^{(\ell)} + \sum_{j\in\mathcal{N}_i\setminus\{i\}} w_{ij}x_j^{(\ell)}, \tag{2-10}$$

where the weights $w_{ij} \geq 0$ quantify the influence of neighbor j on node i, and satisfy the convexity condition $\sum_{j\in\mathcal{N}_i\setminus\{i\}} w_{ij} + w_{ii} = 1$.

Stacking the node states into a global vector as $\mathbf{x}^{(\ell)} = \begin{bmatrix} x_1^{(\ell)} & \cdots & x_N^{(\ell)} \end{bmatrix}^\top \in \mathbb{R}^N$, the update rule becomes

$$\mathbf{x}^{(\ell+1)} = \mathbf{W}\mathbf{x}^{(\ell)}, \tag{2-11}$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$ is the consensus weight matrix with entries $[\mathbf{W}]_{ij} = w_{ij}$ [19].

To achieve average consensus, i.e., convergence of all node values to the global initial average $\bar{x}^{(0)} = \frac{1}{N}\sum_{i=1}^N x_i^{(0)}$, the matrix $\mathbf{W}$ must be doubly stochastic, satisfying $\mathbf{W}\mathbf{1} = \mathbf{1}$ and $\mathbf{W}^\top\mathbf{1} = \mathbf{1}$. These conditions ensure that both the sum and average of node values remain invariant across iterations, i.e., $\bar{x}^{(\ell+1)} = \bar{x}^{(\ell)} = \bar{x}^{(0)}$, and that no bias is introduced into the final consensus value.

Provided the communication graph $\mathcal{G}$ is connected and the weight matrix $\mathbf{W}$ respects symmetry or balance conditions, the iteration converges asymptotically to the true average at all nodes:

$$\lim_{\ell\to\infty} x_i^{(\ell)} = \bar{x}^{(0)}, \quad \forall i \in \mathcal{V}.$$

Laplacian-Based Formulation

An equivalent and particularly insightful formulation of consensus dynamics leverages the graph Laplacian matrix $\mathbf{L}$, transforming the update rule into

$$\mathbf{x}^{(\ell+1)} = \mathbf{x}^{(\ell)} - \epsilon\mathbf{L}\mathbf{x}^{(\ell)} = (\mathbf{I} - \epsilon\mathbf{L})\mathbf{x}^{(\ell)}, \tag{2-12}$$

where $\epsilon > 0$ is a fixed step size that controls the aggressiveness of updates. This Laplacian-based form is not just a restatement of the general weight matrix $\mathbf{W}$ from the previous subsection; it represents a specific parameterization where $\mathbf{W} = \mathbf{I} - \epsilon\mathbf{L}$. This formulation holds for unweighted graphs or symmetric weighting schemes that render $\mathbf{L}$ compatible with a doubly stochastic structure. The key advantage of this view is that it connects consensus directly to spectral graph theory, offering tools for analyzing network connectivity and convergence without explicitly designing weights for every edge. The Laplacian inherently represents the network's structure, making it particularly advantageous for theoretical analysis and optimization in extensive sensor networks.

One of the most insightful elements of this formulation is its interpretation as a gradient descent algorithm that minimizes a quadratic *disagreement function* (also referred to as the

graph Dirichlet form or consensus cost function) [4], which is defined as

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{L}\mathbf{x} = \frac{1}{2}\sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2. \tag{2-13}$$

Here, $\phi(\mathbf{x})$ quantifies the total disagreement across the network by summing squared differences in values between connected nodes (edges in $\mathcal{E}$). A high value of $\phi(\mathbf{x})$ indicates significant discrepancies among neighboring nodes, while $\phi(\mathbf{x}) = 0$ occurs when nodes have identical values (perfect consensus, where $\mathbf{x} = c\mathbf{1}$ for some constant c). The update rule in Eq. (2-12) performs discrete-time gradient descent on this disagreement function, since the gradient is $\nabla\phi(\mathbf{x}) = \mathbf{L}\mathbf{x}$, and subtracting a scaled version (via ϵ) moves the state vector toward the minimum. This optimization perspective explains consensus emergence: each iteration reduces the differences between neighbors, propagating agreement through the graph until reaching $\phi(\mathbf{x})$'s minimum. For stability and convergence, the step size ϵ must be chosen carefully within the range

$$0 < \epsilon < \frac{2}{\lambda_{\max}(\mathbf{L})},$$

where $\lambda_{\max}(\mathbf{L})$ is the largest eigenvalue of $\mathbf{L}$ to prevent overshooting or divergence [12].

2-3-3 Convergence Analysis

Building on the weight-based and Laplacian formulations presented earlier, the conditions under which consensus is achieved and the factors influencing the speed are examined.

Conditions for Convergence

The average consensus is achieved through a combination of topological and algebraic characteristics that enable information to spread throughout the network. For an undirected graph and the linear iteration of Equation Eq. (2-11), convergence to the true average is guaranteed if the following conditions hold:

1. **Connectivity:** The communication graph $\mathcal{G}$ is connected.
2. **Row-stochasticity:** $\mathbf{W}\mathbf{1} = \mathbf{1}$, so every update is a convex combination of the current neighbor values.
3. **Column-stochasticity:** $\mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$. For undirected graphs with symmetric weights, column-stochasticity follows automatically from item 2, so items 2 and 3 imply that $\mathbf{W}$ is doubly stochastic.
4. **Primitivity:** $\mathbf{W}$ is irreducible and aperiodic, typically ensured by assigning a positive self-weight $w_{ii} > 0$. This rules out oscillatory behavior and guarantees the existence of a unique stationary vector.

Under these conditions we have,

$$\lim_{\ell\to\infty} \mathbf{x}^{(\ell)} = \bar{x}^{(0)}\,\mathbf{1}, \qquad \bar{x}^{(0)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(0)},$$

i.e., all nodes asymptotically agree on the initial average.

(Equivalently, one can state that the spectral radius $\rho\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) < 1$; this condition is automatically satisfied when the four points above hold.)

Convergence Rate

The convergence rate, i.e., how quickly nodes reach agreement, is governed by the spectral properties of the consensus matrix $\mathbf{W}$ (introduced in Section 2-3-2), whether defined directly or through a Laplacian-based construction. Let $\bar{x}^{(0)}$ be the initial network-wide average. Defining the consensus error as $\mathbf{e}^{(\ell)} = \mathbf{x}^{(\ell)} - \bar{x}^{(0)}\mathbf{1}$, the update rule becomes

$$\mathbf{e}^{(\ell+1)} = \mathbf{W}\mathbf{e}^{(\ell)}, \tag{2-14}$$

which shows that error evolves linearly via repeated applications of $\mathbf{W}$. Due to the average-invariant property, the consensus component $\bar{x}^{(0)}\mathbf{1}$ remains fixed and lies in the 1-eigenspace of $\mathbf{W}$. Consequently, the error vector $\mathbf{e}^{(\ell)}$ is orthogonal to $\mathbf{1}$, and convergence takes place entirely within the $(N-1)$-dimensional subspace capturing disagreement across nodes.

Let $1 = \lambda_1(\mathbf{W}) \geq \lambda_2(\mathbf{W}) \geq \cdots \geq \lambda_N(\mathbf{W})$ denote the eigenvalues of $\mathbf{W}$. The dominant eigenvalue $\lambda_1 = 1$ corresponds to the invariant consensus mode, while the remaining eigenvalues govern the contraction of the error dynamics. The convergence factor is defined as

$$\rho(\mathbf{W}) = \max\left\{|\lambda_2(\mathbf{W})|, |\lambda_N(\mathbf{W})|\right\}, \tag{2-15}$$

which is the essential spectral radius of $\mathbf{W}$ restricted to the subspace orthogonal to $\mathbf{1}$. The consensus error norm satisfies the geometric bound

$$\left\|\mathbf{x}^{(\ell)} - \bar{x}^{(0)}\mathbf{1}\right\| = \left\|\mathbf{e}^{(\ell)}\right\| \leq \rho(\mathbf{W})^\ell \left\|\mathbf{e}^{(0)}\right\|, \tag{2-16}$$

implying that smaller values of $\rho(\mathbf{W})$ lead to faster convergence. For the Laplacian-based update $\mathbf{W} = \mathbf{I} - \epsilon\mathbf{L}$, where the step size ϵ is chosen to ensure convergence, the error dynamics become

$$\mathbf{e}^{(\ell+1)} = (\mathbf{I} - \epsilon\mathbf{L})\mathbf{e}^{(\ell)}, \tag{2-17}$$

and the eigenvalues of $\mathbf{W}$ are $\mu_i = 1 - \epsilon\lambda_i(\mathbf{L})$, where $0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L})$. The subdominant eigenvalues are thus $\mu_i = 1 - \epsilon\lambda_i(\mathbf{L})$ for $i \geq 2$, and the convergence factor is $\rho(\mathbf{W}) = \max_{i=2,\ldots,N} |1 - \epsilon\lambda_i(\mathbf{L})|$. To minimize this, the step size ϵ should be chosen to equalize the magnitudes of the slowest and fastest contracting modes, leading to the smallest possible convergence factor. This yields the optimal step size

$$\epsilon^* = \frac{2}{\lambda_2(\mathbf{L}) + \lambda_N(\mathbf{L})}, \tag{2-18}$$

resulting in the best-case convergence factor $\rho^* = \frac{\lambda_N(\mathbf{L}) - \lambda_2(\mathbf{L})}{\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})}$. This expression shows how $\lambda_2(\mathbf{L})$ and $\lambda_N(\mathbf{L})$ influence convergence behavior. Faster convergence occurs when λ_2 is large, indicating strong connectivity, and λ_N is small, reflecting a tighter spectral range. These conditions are crucial in WSNs where limited bandwidth and energy demand rapid agreement for efficient estimation.

2-3-4 Extensions for Distributed Estimation

In distributed Kalman filtering, consensus operations extend to matrix-valued and time-varying quantities. One extension is matrix consensus, where nodes reach agreement on local covariance estimates using

$$\mathbf{P}_i^{(\ell+1)} = \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{P}_j^{(\ell)}, \tag{2-19}$$

which preserves positive semidefiniteness when the weights $w_{ij} \geq 0$.

These extensions constitute the core primitives of distributed Kalman filters: they enable state mean fusion, covariance agreement (including covariance–intersection variants [20]), and information-form updates, all with purely local messages. By tuning the number of consensus rounds per time step, a designer can trade estimation accuracy for communication and energy cost, matching the constraints of a given WSN.

2-4 Centralized Kalman Filter

The KF is a cornerstone algorithm for recursive state estimation in noisy dynamic systems, providing optimal performance under linear and Gaussian assumptions. In a centralized setting, all measurements $\{\mathbf{z}_{i,k}\}_{i=1}^N$ are aggregated at a FC, which applies a standard KF to compute the Minimum Mean Squared Estimate (MMSE) estimate of the global system state $\mathbf{x}_k$. Although optimal in theory, this approach incurs significant communication overhead, high energy consumption, and vulnerability to single-point failures.

Building on the system model introduced in Section 2-1, the Centralized Kalman Filter (CKF) aggregates local measurements into a global vector $\mathbf{Z}_k = [\mathbf{z}_{1,k}^\top, \mathbf{z}_{2,k}^\top, \ldots, \mathbf{z}_{N,k}^\top]^\top \in \mathbb{R}^m$, where $m = \sum_{i=1}^N m_i$, together with the stacked measurement matrix $\mathbf{H}_k = [\mathbf{H}_{1,k}^\top, \mathbf{H}_{2,k}^\top, \ldots, \mathbf{H}_{N,k}^\top]^\top \in \mathbb{R}^{m \times n}$, and a block-diagonal noise covariance matrix $\mathbf{R}_k = \operatorname{diag}(\mathbf{R}_{1,k}, \ldots, \mathbf{R}_{N,k}) \in \mathbb{R}^{m \times m}$. This structure assumes that measurement noises $\boldsymbol{\nu}_{i,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{i,k})$ are temporally uncorrelated and mutually uncorrelated across sensor nodes, i.e., $\mathbb{E}[\boldsymbol{\nu}_{i,k}\boldsymbol{\nu}_{j,k}^\top] = \mathbf{0}$ for all $i \neq j$. The resulting global observation model is

$$\mathbf{Z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k), \tag{2-20}$$

with $\mathbf{v}_k = [\boldsymbol{\nu}_{1,k}^\top, \ldots, \boldsymbol{\nu}_{N,k}^\top]^\top$, which enables the joint processing of all sensor data

The CKF operates in two recursive stages: the prediction step advances the prior estimate and covariance using the state dynamics:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}_{k-1} \hat{\mathbf{x}}_{k-1|k-1}, \tag{2-21}$$

$$\mathbf{P}_{k|k-1} = \mathbf{A}_{k-1} \mathbf{P}_{k-1|k-1} \mathbf{A}_{k-1}^\top + \mathbf{Q}_{k-1}. \tag{2-22}$$

The update step then assimilates $\mathbf{Z}_k$ to refine the estimate as follows:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^\top \left(\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^\top + \mathbf{R}_k \right)^{-1}, \tag{2-23}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \left(\mathbf{Z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1} \right), \tag{2-24}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k|k-1}. \tag{2-25}$$

The Kalman gain $\mathbf{K}_k$ optimally balances the prediction uncertainty against the measurement reliability, weighting the innovation (the residual between the observed and predicted measurements) to minimize the posterior variance. The updated covariance $\mathbf{P}_{k|k}$ quantifies the remaining uncertainty after data assimilation.

Under the linear Gaussian framework, the CKF is the conditional mean $\mathbb{E}[\mathbf{x}_k \mid \mathbf{Z}_{1:k}]$, attaining the lowest possible Mean Squared Error (MSE). However, the CKF serves as a performance benchmark for distributed methods, despite its practical limitations in large-scale networks (as discussed in Section 2-1). These motivate alternatives where nodes perform local prediction and updates and leverage consensus mechanisms (detailed in Section 2-3) for neighbor-based fusion.

Chapter 3

Robust Estimation under Outliers

This chapter builds on limitations discussed in Chapter 1 and foundations laid in Chapter 2, addressing outlier contamination in distributed state estimation. In real-world Wireless Sensor Networks (WSNs), measurements can be corrupted by sensor faults, interference, or attacks, leading to extreme deviations that cannot be handled by standard Gaussian-based estimators.

Two representative robust estimation strategies are reviewed, and their practical limitations in terms of scalability and computational cost are discussed. A distributed estimation framework incorporating a smooth log-cosh loss is then introduced to address those limitations. Moreover, the log-cosh loss is compared with other robust loss functions in the literature, justifying its use in the proposed work.

3-1 Motivation

Traditional state estimation methods, including centralized and distributed Kalman Filter (KF), assume Gaussian noise. While optimal under this model, such filters are highly sensitive to large outliers, as the squared residual $\frac{1}{2}r^2$, disproportionately amplifies their influence. In distributed systems, this influence can propagate through consensus updates, thereby distorting the state estimates across nodes. This vulnerability affects WSNs deployments, where noisy channels, hardware issues, or adversarial injection can introduce abnormal observations. These outliers can cause estimation bias, loss of consensus, or divergence, impacting decision-making in tracking, monitoring, or control tasks.

To mitigate the impact of anomalous data, robust estimation frameworks employ loss functions that saturate for large residuals. Examples include the Huber, Tukey, and log-cosh losses, which attenuate the influence of extreme deviations while remaining sensitive to nominal noise. Motivated by this, the current chapter focuses on applying robust loss functions within a distributed filter framework. Special attention is given to the log-cosh penalty, whose convexity and smoothness make it particularly attractive for gradient-based consensus methods like Exact First Order Algorithm (EXTRA) [16]. Through this integration, the aim is

to achieve distributed estimation that is both accurate in clean conditions and robust under contamination.

3-2 Existing Robust Estimation Frameworks

3-2-1 Robust Consensus Filter in the Presence of Student-t Distributed Measurement Noise

The authors in [13] use the Student-t distribution as the measurement noise model in their Robust Consensus Nonlinear Information Filter (RCNIF). This choice addresses limitations of traditional filters, which are designed under the assumption of Gaussian-distributed measurement noise. Specifically, conventional approaches like the KF and its nonlinear variants (e.g., unscented or cubature filters) assume measurement noise follows a Gaussian distribution, with a symmetric bell curve and rapidly decaying tails. While this works for well-behaved noise, it fails with outlier measurements from sensor malfunctions, environmental disturbances, or anomalies. Such outliers create heavy-tailed noise, where extreme values occur more frequently than predicted by Gaussian models, potentially causing filter divergence or estimation errors. The Student-t distribution, with heavier tails, provides a more flexible model for such scenarios, accommodating rare deviations without requiring prior knowledge of their occurrence. This makes it suitable for WSNs, where sensors operate independently and may experience varying noise corruption.

The Student-t distribution use is justified by its ability to interpolate between extreme distributions via its Degree of Freedom (DoF) parameter, ν. When $\nu = 1$, it reduces to the Cauchy distribution, which has extremely heavy tails and no defined mean or variance, modeling severe outliers. As ν increases, the tails lighten, and as $\nu \to \infty$, it converges to a Gaussian distribution, aligning with nominal conditions. This adaptability allows the RCNIF to handle noise behaviors, from mild deviations to significant outliers, within a single framework.

Unlike Gaussian models, which assign negligible probability to extreme values, the Student-t distribution's heavier tails ensure that outliers are not dismissed as impossible but are instead integrated into the estimation process with appropriate weighting. The measurement noise is modeled as

$$\mathbf{r}_{i,k} \sim \mathrm{St}(0, (\mathbf{\Lambda}_{i,k})^{-1}, \nu_{i,k}) \tag{3-1}$$

for each sensor node i, where

$$\mathbf{z}_{i,k} = h_i(\mathbf{x}_k) + \mathbf{r}_{i,k}. \tag{3-2}$$

The RCNIF relies on the representation of Student-t distributions as a mixture model, which simplifies Bayesian inference and filtering computations. Specifically, a Student-t random variable can be expressed as a Gaussian distribution scaled by a Gamma-distributed auxiliary variable. For a measurement $\mathbf{z}_{i,k}$ at node i, the model is:

$$\mathbf{z}_{i,k}|\mathbf{x}_k, \lambda_{i,k} \sim \mathcal{N}\big(h_i(\mathbf{x}_k), (\lambda_{i,k}\Lambda_{i,k})^{-1}\big), \tag{3-3}$$

$$\lambda_{i,k} \sim \mathrm{Gamma}(v_{i,k}/2, v_{i,k}/2). \tag{3-4}$$

Here, $\lambda_{i,k}$ acts as a latent scaling factor adjusting the Gaussian component variance. When $\lambda_{i,k}$ is large, the covariance $(\lambda_{i,k}\mathbf{\Lambda}_{i,k})^{-1}$ shrinks, resembling Gaussian noise; when $\lambda_{i,k}$ is small,

the covariance expands to capture outliers. The marginal distribution of $\mathbf{z}_{i,k}$ (integrating out $\lambda_{i,k}$) yields the Student-t form:

$$\begin{aligned} p(\mathbf{z}_{i,k}|\mathbf{x}_k) &= \int \mathcal{N}\big(\mathbf{z}_{i,k}; h_i(\mathbf{x}_k), (\lambda_{i,k}\mathbf{\Lambda}_{i,k})^{-1}\big) \cdot \text{Gamma}\big(\lambda_{i,k}; v_{i,k}/2, v_{i,k}/2\big) d\lambda_{i,k} \\ &= \text{St}\big(\mathbf{z}_{i,k}; h_i(\mathbf{x}_k), (\mathbf{\Lambda}_{i,k})^{-1}, v_{i,k}\big). \end{aligned} \tag{3-5}$$

This hierarchical structure is computationally advantageous because it transforms the non-Gaussian Student-t problem into a conditionally Gaussian one, making it amenable to techniques like the information filter once $\lambda_{i,k}$ is estimated.

The mixture model introduces unknown parameters that must be estimated alongside the state $\mathbf{x}_k$: the precision matrix $\mathbf{\Lambda}_{i,k}$ and the DoF $v_{i,k}$. These parameters define the shape and spread of the noise distribution, yet are typically unavailable a priori in real-world applications. The precision matrix $\mathbf{\Lambda}_{i,k}$ for each sensor node i at time k models the measurement noise within the Student-t framework. The precision matrix is the inverse of the scale matrix $(\mathbf{\Lambda}_{i,k})^{-1}$ in the Student-t distribution, governing the spread and correlation of the noise $\mathbf{r}_{i,k}$ in the measurement model. The authors model $\mathbf{\Lambda}_{i,k}$ as a random variable following a Wishart distribution:

$$\mathbf{\Lambda}_{i,k} \sim \mathcal{W}(v_{i,k}, \mathbf{V}_{i,k}), \tag{3-6}$$

where $v_{i,k}$ represents the DoF and $\mathbf{V}_{i,k}$ is a symmetric, positive definite scale matrix. This distribution choice allows $\mathbf{\Lambda}_{i,k}$ to adapt dynamically to the data, capturing the uncertainty and variability in the noise characteristics, especially in the presence of outliers.

The Wishart distribution is the conjugate prior for the precision matrix of a multivariate Gaussian distribution, aligning with the mixture representation of the Student-t model: $\mathbf{z}_{i,k}|\mathbf{x}_k, \lambda_{i,k} \sim \mathcal{N}(h_i(\mathbf{x}_k), (\lambda_{i,k}\mathbf{\Lambda}_{i,k})^{-1})$. Conjugacy ensures the posterior distribution of $\mathbf{\Lambda}_{i,k}$ remains Wishart after incorporating data, simplifying Bayesian updates. It suits positive definite matrices like $\mathbf{\Lambda}_{i,k}$, ensuring the precision matrix stays valid during estimation. Parameters $v_{i,k}$ and $\mathbf{V}_{i,k}$ provide flexibility: $v_{i,k}$ controls prior confidence (higher values indicate tighter concentration around expected precision), while $\mathbf{V}_{i,k}$ scales the expected precision matrix. A dynamic model for these parameters is introduced: $v_k^i = \rho v_{k-1}^i$ and $\mathbf{V}_k^i = \mathbf{B}\mathbf{V}_{k-1}^i\mathbf{B}^T$ with $\mathbf{B} = 1/\sqrt{\rho}$ and discount factor $0 < \rho \leq 1$ to model temporal evolution, balancing adaptability with stability. This enhances the RCNIF's robustness by enabling noise structure adaptation, outperforming fixed or Gaussian-based noise models.

Similarly, $v_{i,k}$ is modeled as a Gamma distribution, $v_{i,k} \sim \text{Gamma}(\alpha_{i,k}, \beta_{i,k})$, reflecting its positive nature and allowing flexibility in tail heaviness.

Variational Bayesian (VB) inference tackles the challenge of approximating the joint posterior

$$p(\mathbf{x}_k, \lambda_k, \mathbf{\Lambda}_k, v_k|\mathcal{Z}_k), \tag{3-7}$$

which is intractable due to the coupling of the state $\mathbf{x}_k$ and noise parameters $\lambda_{i,k}$, $\mathbf{\Lambda}_{i,k}$, and $v_{i,k}$ in a nonlinear, non-Gaussian system with measurement outliers. To simplify this, VB assumes a factorized form for the approximate posterior:

$$q(\mathbf{x}_k)q(\lambda_k)q(\mathbf{\Lambda}_k)q(v_k), \tag{3-8}$$

where:

- $q(\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_k; \hat{\mathbf{x}}_k, \mathbf{P}_k)$: Gaussian for the state.
- $q(\lambda_k) = \prod_{i \in S} q_i(\lambda_{i,k})$, with $q_i(\lambda_{i,k}) = \text{Gamma}(\lambda_{i,k}; a_{i,k}, b_{i,k})$: Gamma for each $\lambda_{i,k}$.
- $q(\mathbf{\Lambda}_k) = \prod_{i \in S} q_i(\mathbf{\Lambda}_{i,k})$, with $q_i(\mathbf{\Lambda}_{i,k}) = \mathcal{W}(\mathbf{\Lambda}_{i,k}; v_{i,k}, \mathbf{V}_{i,k})$: Wishart for each $\mathbf{\Lambda}_{i,k}$.
- $q(v_k) = \prod_{i \in S} q_i(v_{i,k})$, with $q_i(v_{i,k}) = \text{Gamma}(v_{i,k}; \alpha_{i,k}, \beta_{i,k})$: Gamma for each $v_{i,k}$.

This factorization imposes independence among variables in the approximation, even though they are coupled in the true posterior. The optimal q-distributions are determined by minimizing the Kullback-Leibler (KL) divergence between this approximation and the true posterior, a process equivalent to maximizing the Evidence Lower Bound (ELBO), defined as:

$$\text{ELBO} = \int q(\mathbf{x}_k) q(\lambda_k) q(\mathbf{\Lambda}_k) q(v_k) \ln \left(\frac{p(\mathbf{x}_k, \lambda_k, \mathbf{\Lambda}_k, v_k, \mathcal{Z}_k)}{q(\mathbf{x}_k) q(\lambda_k) q(\mathbf{\Lambda}_k) q(v_k)} \right) d\mathbf{x}_k d\lambda_k d\mathbf{\Lambda}_k dv_k. \tag{3-9}$$

This transformation converts an intractable integration into a practical optimization task, leveraging conjugate properties of chosen distributions (e.g., Gaussian with Gaussian, Gamma with Gaussian precision) to ensure each q-distribution maintains its parametric form. VB uses coordinate ascent, iteratively updating each factor $q(\cdot)$ while holding others fixed. The updates derive from the joint distribution's log-likelihood, averaged over other variables' current q-distributions.

Beyond local estimation, VB's role extends to the distributed consensus framework of the RCNIF, where each sensor node performs its own VB updates using local measurements $\mathbf{z}_{i,k}$, producing factorized posteriors that are then fused across the network via the Hybrid Consensus on Measurement and Information (HCMCI) method proposed in section [21]. The prediction step propagates the state and parameters forward, setting priors like

$$q(\mathbf{x}_k | \mathcal{Z}_{k-1}) = \mathcal{N}(\hat{\mathbf{x}}_{k|k-1}, P_{k|k-1}), \tag{3-10}$$

while the update step refines these with new data, integrating likelihood contributions (e.g., $\delta q_k^{i,t}, \delta \Omega_k^{i,t}$) into consensus. VB's iterative refinement ensures each node's estimate adapts to local noise characteristics, and its information-form updates align with the information filter's fusion-friendly structure, facilitating averaging of priors and likelihoods across nodes. This dual role, local posterior approximation and distributed consensus support, shows VB's importance, enabling the RCNIF to achieve robust state estimation in distributed sensor networks under challenging outlier conditions, outperforming traditional Gaussian-based filters.

3-2-2 Covariance Inflation

The paper [15] introduces a novel approach to online filtering in probabilistic State Space Models (SSMs) robust to outliers and misspecified measurement models. Traditional Kalman filtering methods, such as KF, Extended Kalman Filter (EKF), and Ensemble Kalman Filter (EnKF), assume Gaussian measurement models and linear dynamics, making them sensitive to outliers and model misspecification. This sensitivity can degrade performance in real-world applications, including tracking, time-series forecasting, and neural network training, where data anomalies and measurement errors are common. To address these challenges,

Algorithm 1 WoLF predict and update step

Require: $\mathbf{A}$, $\mathbf{Q}$ *// predict step*

$\hat{\mathbf{x}}_{k|k-1} \leftarrow \mathbf{A}\hat{\mathbf{x}}_{k-1|k-1}$

$\boldsymbol{P}_{k|k-1} \leftarrow \mathbf{A}\boldsymbol{P}_{k-1|k-1}\mathbf{A}^\top + \mathbf{Q}$

Require: $\mathbf{y}_k$, $\mathbf{H}$, $\mathbf{R}$ *// update step*

$\hat{\mathbf{y}}_k \leftarrow \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}$

$w_k \leftarrow W(\mathbf{y}_k, \hat{\mathbf{y}}_k)$

$\boldsymbol{P}^{-1}_{k|k} \leftarrow \boldsymbol{P}^{-1}_{k|k-1} + w_k^2\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}$

$\mathbf{K}_k \leftarrow w_k^2\boldsymbol{P}_{k|k}\mathbf{H}^\top\mathbf{R}^{-1}$

$\hat{\mathbf{x}}_{k|k} \leftarrow \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \hat{\mathbf{y}}_k)$

the authors proposed Weighted Observation Likelihood Filter (WoLF), which integrates generalized Bayesian inference with traditional filtering techniques. Instead of a standard log-likelihood, WoLF employs a weighted loss function to down-weight outliers while retaining closed-form Gaussian posterior updates, ensuring computational efficiency, though it remains a centralized rather than distributed approach.

Generalized Bayesian inference extends traditional Bayesian methods by replacing log-likelihood with a flexible loss function, providing robust statistical modeling for model misspecification or outliers. Unlike classical Bayes theorem, which updates prior beliefs using data likelihood under a specific model, generalized Bayes uses a loss function $\ell(\boldsymbol{\theta}, \mathbf{y})$ that quantifies discrepancy between model parameters $\boldsymbol{\theta}$ and observations $\mathbf{y}$. The generalized posterior is then defined as:

$$q(\boldsymbol{\theta}|\mathbf{y}) \propto \exp(-\ell(\boldsymbol{\theta}, \mathbf{y}))p(\boldsymbol{\theta}), \tag{3-11}$$

where $p(\boldsymbol{\theta})$ is the prior distribution. This formulation allows for tailoring inference to specific needs, such as down-weighting outliers or handling non-standard distributions, while retaining probabilistic interpretation. By generalizing the update rule, this approach provides a sound and feasible alternative to traditional methods.

The WoLF builds on this framework as an adaptation of the KF. In the standard KF, the update step assumes a Gaussian likelihood, making it highly sensitive to extreme observations. WoLF addresses this limitation by incorporating a loss function inspired by generalized Bayes, replacing the traditional negative log-likelihood with a weighted version:

$$\ell_t(\boldsymbol{\theta}_t) = -W^2(\mathbf{y}_t, \hat{\mathbf{y}}_t)\log q(\mathbf{y}_t|\boldsymbol{\theta}_t). \tag{3-12}$$

Here, $W(\mathbf{y}_t, \hat{\mathbf{y}}_t)$ is a weighting function that adjusts each observation $\mathbf{y}_t$ based on its deviation from the predicted observation $\hat{\mathbf{y}}_t$. This weighting function reduces the impact of outliers while preserving the closed-form Gaussian posterior update characteristics of the KF. Consequently, the WoLF Algorithm 1 ensures computational efficiency, matching the standard KF, unlike other robust methods.

The flexibility of WoLF lies in its choice of weighting functions, which allow for various outlier-handling strategies. Notable variants include the Inverse Multi-Quadratic (WoLF-IMQ), the Mahalanobis Distance (WoLF-MD), and the Threshold Mahalanobis Distance (WoLF-TMD), each of which provides different mechanisms for mitigating the impact of extreme observations.

The IMQ weight is defined as:

$$W(\mathbf{y}_t, \hat{\mathbf{y}}_t) = \left(1 + \frac{\|\mathbf{y}_t - \hat{\mathbf{y}}_t\|_2^2}{c^2}\right)^{-1/2}, \tag{3-13}$$

which employs the Euclidean (ℓ_2) distance between the observation $\mathbf{y}_t$ and its prediction $\hat{\mathbf{y}}_t$, scaled by a soft threshold c. This function smoothly reduces the influence of outliers as their distance from the prediction increases, providing a compensation-based approach that retains some information from extreme observations.

In contrast, the Mahalanobis distance weight is given by:

$$W(\mathbf{y}_t, \hat{\mathbf{y}}_t) = \left(1 + \frac{\|\mathbf{R}_t^{-1/2}(\mathbf{y}_t - \hat{\mathbf{y}}_t)\|_2^2}{c^2}\right)^{-1/2}, \tag{3-14}$$

which incorporates the measurement covariance $\mathbf{R}_t$, adjusting the distance metric to account for the noise structure and correlations in the data. This makes it more adaptive to the statistical properties of the measurement process, offering refined robustness by aligning the weight with the expected variability of observations.

Both weighting mechanisms ensure outliers exert a bounded influence, which has been theoretically proven. This makes WoLF computationally efficient for real-world filtering applications. The algorithm follows the prediction-update structure of the KF. In the prediction step, the state estimate propagates using the dynamics model. In the update step, the estimate is refined by incorporating the weighted observation likelihood. Theoretical robustness is ensured through a bounded Posterior Influence Function (PIF), which limits the impact of extreme outliers on the posterior distribution.

3-2-3 Limitations of Existing Frameworks

Existing robust filtering methods, such as the RCNIF and WoLF, address outlier-prone measurements in distributed and centralized settings but have structural limitations that constrain their efficiency, scalability, and reliability. These limitations, rooted in computational complexity and implementation challenges, motivate the development of a convex-loss-based framework that prioritizes simplicity while maintaining robustness.

The RCNIF employs a Student-t noise model with VB inference for heavy tail outliers. In linear state-space models with affine dynamics and measurements, conjugacy enables closed-form VB updates resembling an Expectation-Maximization (EM) procedure: an E-step computes the expected precision $\bar{\lambda}_k \triangleq \mathbb{E}_{q(\lambda_k)}[\lambda_k]$, followed by an M-step performing a weighted Kalman update with effective covariance

$$\mathbf{R}_k^{\text{eff}} = \left(\bar{\lambda}_k \cdot \mathbb{E}_{q(\mathbf{\Lambda}_k)}[\mathbf{\Lambda}_k]\right)^{-1}.$$

The method requires updating three latent variables per node and time step via multiple coordinate-ascent sweeps, doubling the computation and memory demands compared to a standard KF. In distributed networks, consensus via the HCMCI requires sharing latent statistics with state updates, increasing the bandwidth and synchronization latency.

Hyperparameter sensitivity affects robustness through the initial degrees of freedom v_0, Wishart scale $\mathbf{V}_0$, and discount factor ρ, without principled tuning methods. Poor choices lead to data mishandling, whereas the non-convex ELBO risks suboptimal convergence. The mean-field factorization ignores posterior couplings and often underestimates covariances.

The WoLF adapts the KF via generalized Bayesian inference, using a weighted log-likelihood, where W down-weights outliers based on the prediction distance. While matching the KF efficiency, its scalar weighting affects all measurement dimensions uniformly, suppressing reliable channels when outliers occur. The fixed robustness threshold lacks adaptability to time-varying statistics. As a centralized method, WoLF requires extensions for distributed fusion.

These limitations motivate the scaled log-cosh framework, which eliminates latent variables and VB iterations, providing a strictly convex C^∞ objective with a unique minimizer tunable by α. It enables lightweight consensus via EXTRA, and rigorous stability analysis, achieving robustness with reduced overhead.

3-3 Proposed Strategy: Log-Cosh Loss-Based Robust Estimation

A method based on the log-cosh loss is proposed, which builds on the existing robust strategy WoLF discussed in Section 3-2-2, which uses generalized loss functions for enhanced robustness. In particular, a log-cosh penalty is used instead of the traditional quadratic term in the Maximum A Posteriori (MAP) estimation objective, producing a robust, smooth, and convex formulation. This method offers a well-balanced compromise: it reduces the impact of large outliers while maintaining responsiveness to small residuals, guaranteeing precision under nominal conditions (i.e., no outliers). Because of these characteristics, the log-cosh loss is especially well-suited for reliable distributed estimation in outlier-prone environments.

For each scalar $x \in \mathbb{R}$, the log-cosh loss is defined as follows:

$$L(x) = \log\cosh(x). \tag{3-15}$$

For small arguments, it acts approximately quadratically ($L(x) \approx \frac{1}{2}x^2$), whereas for large arguments, it behaves linearly ($L(x) \approx |x| - \log 2$). As such, it smoothly switches between squared and absolute value losses, making it ideal for robust regression when heavy-tailed noise is present.

The loss function is applied to the components of the whitened residual vector, defined as

$$\tilde{r}_{i,k,l}(\mathbf{x}_k) = \left[\mathbf{R}_{i,k}^{-1/2}\left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right)\right]_l, \quad l = 1, \ldots, m_i. \tag{3-16}$$

This corresponds to the scaled innovation terms that normalize each residual component by its local noise variance. Here, $\mathbf{R}_{i,k} \in \mathbb{R}^{m_i \times m_i}$ is the measurement noise covariance matrix linked to the nominal Gaussian approximation of the true heavy-tailed noise model (unimodal and symmetric). $\mathbf{R}_{i,k}$ is used to describe the whitening transformation, which standardizes residuals across sensors and maintains the Mahalanobis structure of the cost, even when the actual noise may not be Gaussian. This allows robust estimation while guaranteeing compatibility with traditional KF updates (see Eq. (3-18) and Eq. (3-24)).

A scaling parameter $\alpha > 0$ is introduced to regulate the robustness level, and each component is subjected to the log-cosh loss as $L_{i,k}(\alpha \tilde{r}_{i,k,l}(\mathbf{x}_k))$. A loss function that closely mimics the Gaussian negative log-likelihood is produced by smaller values of α, which provides poor robustness but strong sensitivity to nominal data. Conversely, larger values of α saturate the loss for moderate to large residuals, suppressing their influence. This enhances the robustness at the expense of decreased sensitivity to small deviations. A normalization factor $1/\alpha^2$ is provided to guarantee that when residuals are small, the resulting cost closely resembles the traditional quadratic loss. At time step k, this results in the local robust objective function as follows:

$$L_{i,k}(\mathbf{x}_k) = \frac{1}{\alpha^2} \sum_{l=1}^{m_i} \log\cosh\left(\alpha\, \tilde{r}_{i,k,l}(\mathbf{x}_k)\right). \tag{3-17}$$

For small residuals $\tilde{r}_{i,k,l}(\mathbf{x}_k) \approx 0$, $\log\cosh(\alpha\tilde{r}) \approx \frac{1}{2}\alpha^2\tilde{r}^2$, and hence

$$L_{i,k}(\mathbf{x}_k) \approx \frac{1}{2} \sum_{l=1}^{m_i} \tilde{r}_{i,k,l}(\mathbf{x}_k)^2 = \frac{1}{2} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right)^\top \mathbf{R}_{i,k}^{-1} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right), \tag{3-18}$$

which recovers the standard squared Mahalanobis loss.

3-3-1 Log-Cosh in the MAP Estimation Framework

Bayes' theorem is used to incorporate the log-cosh loss into a probabilistic estimation framework. It combines prior knowledge of the state with the likelihood of observations. Specifically, the posterior density $p(\mathbf{x}_k \mid \mathbf{z}_{i,k})$ needs to be computed, which represents the updated belief about the hidden state $\mathbf{x}_k$ after incorporating the measurement $\mathbf{z}_{i,k}$. This posterior is proportional to the product of the prior density $p(\mathbf{x}_k)$ (predictions from previous time steps) and the likelihood $p(\mathbf{z}_{i,k} \mid \mathbf{x}_k)$ (measurement model):

$$p(\mathbf{x}_k \mid \mathbf{z}_{i,k}) \propto p(\mathbf{z}_{i,k} \mid \mathbf{x}_k)\, p(\mathbf{x}_k). \tag{3-19}$$

The MAP estimate identifies the most probable state given the observation:

$$\hat{\mathbf{x}}_{i,k|k} = \arg\max_{\mathbf{x}_k} p(\mathbf{x}_k \mid \mathbf{z}_{i,k}). \tag{3-20}$$

This can be equivalently reformulated as a minimization problem by taking the negative logarithm of the posterior (which preserves the location of the maximum):

$$\hat{\mathbf{x}}_{i,k|k} = \arg\min_{\mathbf{x}_k} -\log p(\mathbf{x}_k \mid \mathbf{z}_{i,k}) = \arg\min_{\mathbf{x}_k} J_{i,k}(\mathbf{x}_k), \tag{3-21}$$

where the cost function $J_{i,k}(\mathbf{x}_k)$ is defined as

$$J_{i,k}(\mathbf{x}_k) \triangleq -\log p(\mathbf{z}_{i,k} \mid \mathbf{x}_k) - \log p(\mathbf{x}_k). \tag{3-22}$$

The first term in Eq. (3-22), also referred to as the negative log-likelihood, is interpreted as the loss that measures the discrepancy between the measured and predicted measurements. The generalized Bayesian framework, as discussed in Section 3-2-2, allows this formulation to be extended by replacing the Gaussian negative log-likelihood with an alternative loss function. This loss function can be chosen to be more robust than the typical quadratic loss used in

the standard KF. In particular, the log-cosh loss can be interpreted as a generalized negative log-likelihood, yielding a convex and smooth objective that enhances robustness.

Now, the estimation of the hidden state $\mathbf{x}_k$ is considered based on the measurement model (see Section 2-1):

$$\mathbf{z}_{i,k} = \mathbf{H}_{i,k}\mathbf{x}_k + \boldsymbol{\nu}_{i,k}, \tag{3-23}$$

where $\boldsymbol{\nu}_{i,k}$ denotes the measurement noise. Under the classical Gaussian assumption, i.e., $\boldsymbol{\nu}_{i,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{i,k})$, the cost function becomes

$$\begin{aligned} J_{i,k}(\mathbf{x}_k) = &\frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1})^\top \mathbf{P}_{i,k|k-1}^{-1}(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1}) \\ &\frac{1}{2}\left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right)^\top \mathbf{R}_{i,k}^{-1}\left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right), \end{aligned} \tag{3-24}$$

which yields the standard KF update.

To improve robustness against outliers, the Gaussian negative log-likelihood term is replaced with a log-cosh-based penalty applied to the whitened residuals (see Eq. (3-16)). This leads to the following cost function:

$$\begin{aligned} J_{i,k}(\mathbf{x}_k) = &\frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1})^\top \mathbf{P}_{i,k|k-1}^{-1}(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1}) \\ &+ \frac{1}{\alpha^2}\sum_{l=1}^{m_i} \log\cosh\left(\alpha\, \tilde{r}_{i,k,l}(\mathbf{x}_k)\right) \end{aligned} \tag{3-25}$$

with the whitened residuals $\tilde{r}_{i,k,l}(\mathbf{x}_k)$ defined in Eq. (3-16).

The objective in Eq. (3-25) replaces the Gaussian negative log-likelihood with a smooth and convex alternative that limits the impact of outliers. The first term retains the standard quadratic prior centered at $\hat{\mathbf{x}}_{i,k|k-1}$, ensuring regularization. The overall cost function $J_{i,k}(\mathbf{x}_k)$ is twice continuously differentiable, and its Hessian is strictly positive definite due to the additive contribution of the prior term $\mathbf{P}_{i,k|k-1}^{-1} \succ 0$. As a result, the objective admits a unique global minimizer, and the posterior density concentrates around this mode.

According to Laplace's approximation [22], the resulting posterior can be locally approximated by a Gaussian distribution centered at the MAP estimate $\hat{\mathbf{x}}_{i,k|k}$. Its covariance is given by the inverse of the Hessian of $J_{i,k}(\mathbf{x}_k)$ evaluated at $\hat{\mathbf{x}}_{i,k|k}$

$$\mathbf{P}_{i,k|k} = \left[\nabla^2 J_{i,k}(\hat{\mathbf{x}}_{i,k|k})\right]^{-1} = \left[\mathbf{P}_{i,k|k-1}^{-1} + \mathbf{H}_{i,k}^\top \mathbf{R}_{i,k}^{-\frac{1}{2}} \mathbf{W}_{i,k}(\hat{\mathbf{x}}_{i,k|k}) \mathbf{R}_{i,k}^{-\frac{1}{2}} \mathbf{H}_{i,k}\right]^{-1}, \tag{3-26}$$

where $\mathbf{W}_{i,k}(\hat{\mathbf{x}}_{i,k|k}) \in \mathbb{R}^{m\times m}$ is a diagonal matrix with entries

$$\left[\mathbf{W}_{i,k}(\hat{\mathbf{x}}_{k|k})\right]_{ll} = \operatorname{sech}^2\left(\alpha \tilde{r}_{i,k,l}(\hat{\mathbf{x}}_{i,k|k})\right), \quad l = 1, \ldots, m. \tag{3-27}$$

This expression generalizes the posterior covariance of the KF by introducing a data-dependent weighting matrix $\mathbf{W}_{i,k}$. The diagonal elements of $\mathbf{W}_{i,k}$ act as an adaptive information scaling. For small residuals, where $\operatorname{sech}^2(\cdot) \approx 1$, the posterior covariance matches the standard KF posterior covariance. For large residuals, where $\operatorname{sech}^2(\cdot) \ll 1$, the contribution of the unreliable observations is downweighted. This mechanism ensures that the posterior uncertainty reflects not only the prior and nominal noise covariances but also the reliability of the observed data. This form allows uncertainty to propagate in the recursive filter updates.

3-3-2 Distributed Implementation via Gradient Consensus

Extending the robust MAP estimator to distributed settings (building on the consensus foundations in Chapter 2 and the system model in Section 2-1), each node $i \in \mathcal{V}$ observes local measurements $\mathbf{z}_{i,k} \in \mathbb{R}^{m_i}$ and seeks to estimate the global state $\mathbf{x}_k$. It does this by combining its local information with that of its neighbors. The sensor network is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and each node communicates only with nodes in its neighborhood $\mathcal{N}_i$.

Assuming that each node first performs a local prediction step (e.g., as in the Centralized Kalman Filter (CKF) described in Section 2-4) to obtain $\hat{\mathbf{x}}_{i,k|k-1}$ and $\mathbf{P}_{i,k|k-1}$, the global MAP estimation problem can be solved in a distributed manner. Each node maintains a local copy $\mathbf{x}_{i,k} \in \mathbb{R}^n$ of the global state $\mathbf{x}_k$ and solves the following constrained optimization problem:

$$\begin{aligned}\min_{\{\mathbf{x}_{i,k}\}} \quad & \sum_{i=1}^{N} \left[\frac{1}{\alpha^2} \sum_{l=1}^{m_i} \log\cosh\left(\alpha \tilde{r}_{i,k,l}(\mathbf{x}_{i,k})\right) \right. \\ & \left. + \frac{1}{2N} \left(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}\right)^\top \mathbf{P}_{i,k|k-1}^{-1} \left(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}\right) \right], \quad \text{s.t. } \mathbf{x}_{i,k} = \mathbf{x}_{j,k},\ \forall (i,j) \in \mathcal{E},\end{aligned} \tag{3-28}$$

where the local whitened residual is defined as

$$\tilde{r}_{i,k,l}(\mathbf{x}_{i,k}) = \left[\mathbf{R}_{i,k}^{-1/2} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_{i,k}\right)\right]_l, \quad l = 1, \ldots, m_i. \tag{3-29}$$

The consensus constraints enforce agreement among neighboring state estimates. Each node minimizes an objective that comprises two components. The first is a robust data fidelity term based on the log-cosh loss, and the second is a Gaussian prior term that incorporates the local prediction $\hat{\mathbf{x}}_{i,k|k-1}$ and its covariance $\mathbf{P}_{i,k|k-1}$.

To solve Eq. (3-28), the EXTRA algorithm proposed in [16] is adopted, which is suitable for smooth convex objectives and supports fixed step-size convergence. The update rule at each node i at iteration ℓ is given by

$$\mathbf{x}_{i,k}^{(\ell+1)} = \mathbf{x}_{i,k}^{(\ell)} + \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_{j,k}^{(\ell)} - \sum_{j \in \mathcal{N}_i} \tilde{w}_{ij} \mathbf{x}_{j,k}^{(\ell-1)} - \epsilon \left[\nabla J_{i,k}(\mathbf{x}_{i,k}^{(\ell)}) - \nabla J_{i,k}(\mathbf{x}_{i,k}^{(\ell-1)})\right], \tag{3-30}$$

where $\epsilon > 0$ is the step size and w_{ij} is the symmetric consensus weight, such as those based on the Metropolis-Hastings rule. The local objective function $J_{i,k}(\mathbf{x}_{i,k})$ is given by

$$J_{i,k}(\mathbf{x}_{i,k}) = \frac{1}{\alpha^2} \sum_{l=1}^{m_i} \log\cosh\left(\alpha \tilde{r}_{i,k,l}(\mathbf{x}_{i,k})\right) + \frac{1}{2N} \left(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}\right)^\top \mathbf{P}_{i,k|k-1}^{-1} \left(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}\right). \tag{3-31}$$

The corresponding gradient is

$$\nabla J_{i,k}(\mathbf{x}_{i,k}) = -\frac{1}{\alpha} \mathbf{H}_{i,k}^\top \mathbf{R}_{i,k}^{-1/2} \boldsymbol{\psi}_i + \frac{1}{N} \mathbf{P}_{i,k|k-1}^{-1} \left(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}\right), \tag{3-32}$$

where $\boldsymbol{\psi}_i \in \mathbb{R}^{m_i}$ has components

$$[\boldsymbol{\psi}_i]_l = \tanh\left(\alpha \tilde{r}_{i,k,l}(\mathbf{x}_{i,k})\right), \quad l = 1, \ldots, m_i. \tag{3-33}$$

After the consensus converges, each node adopts the common estimate $\hat{\mathbf{x}}_{k|k} = \mathbf{x}_{i,k}^{(\ell)}$ as its posterior mean. The corresponding posterior covariance can then be approximated by inverting the global Hessian of the objective, giving

$$\mathbf{P}_{k|k} \approx \left[\sum_{i=1}^{N} \nabla^2 J_{i,k}(\hat{\mathbf{x}}_{k|k}) \right]^{-1}. \tag{3-34}$$

This Hessian can be computed distributively by having each node evaluate its local Hessian and run an average consensus protocol to compute the sum.

This formulation results in a smooth, fully distributed optimization algorithm that is robust to the presence of outliers. The use of EXTRA ensures efficient convergence, making it suitable for resource-constrained sensor networks. Next, this loss function is compared with alternative robust losses.

3-4 Alternative Loss Functions for Robust Estimation

Although log-cosh loss offers a smooth and convex robust alternative to quadratic penalties, many other loss functions reduce outlier influence in estimation tasks. These alternatives limit large residual effects but vary in formulations, convexity, smoothness, and outlier handling, affecting optimization and statistical effectiveness. This section examines key robust loss functions: beginning with convex Huber loss [23, 24], proceeding to non-convex Welsch loss [25], continuing with Tukey's biweight loss [26], followed by logarithmic Least Logarithmic Absolute Difference (LLAD) [27] loss, and concluding with information-theoretic approaches like Maximum Correntropy Criterion (MCC) [28, 29] and its extension, Kernel Risk-Sensitive Loss (KRSL) [30]. This progression shows the shift from bounded-influence designs to aggressive suppression mechanisms and kernel-based metrics, highlighting trade-offs in convexity, parameter sensitivity, and computational complexity.

3-4-1 Huber Loss

The Huber loss [31] function is a well-known robust penalty that serves as a bridge between the squared error and absolute deviation. Initially introduced in the context of M-estimation, it aims to maintain high statistical efficiency under standard conditions while enhancing resistance to outliers. The function is defined in a piecewise manner, applying a quadratic penalty to small residuals and a linear penalty to larger ones.

For a scalar residual r, the Huber loss is given by

$$L_{\text{Huber}}(r;\delta) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le \delta, \\ \delta|r| - \frac{1}{2}\delta^2 & \text{if } |r| > \delta, \end{cases} \tag{3-35}$$

where $\delta > 0$ serves as a tuning parameter that determines the threshold distinguishing inlier from outlier regimes. For residuals within $[-\delta, \delta]$, the loss functions like a squared error. However, for larger residuals, it transitions to a linear penalty, thereby limiting the outlier impact.

The Huber loss is globally convex, ensuring a unique global minima. However, it is only once differentiable (C^1) and lacks second-order smoothness. The discontinuity in the second derivative at $|r| = \delta$ results in a curvature discontinuity, which can hinder the application of second-order optimization techniques.

The influence function, obtained by differentiating L_{Huber}, is

$$\psi_{\text{Huber}}(r;\delta) = \begin{cases} r & \text{if } |r| \leq \delta, \\ \delta \operatorname{sign}(r) & \text{if } |r| > \delta. \end{cases} \tag{3-36}$$

The function is both bounded and monotonic, effectively restricting the gradient's value to δ. This ensures that extreme residuals have a limited impact on the estimation. As a result, the estimator remains resilient to significant errors while maintaining high efficiency in the presence of standard Gaussian noise.

The performance of Huber loss depends on δ, which governs the transition between quadratic and linear regimes. Reducing δ can improve robustness by restricting moderate residuals' impact, although it might underutilize informative data. Larger values reduce robustness as the loss resembles standard quadratic, diminishing its advantage in outlier-contaminated scenarios. Therefore, selecting appropriate δ is essential to balance robustness and efficiency.

The Huber loss provides a trade-off between efficiency and robustness through its convex piecewise formulation. When δ is large, the quadratic region dominates, reducing to squared error. As δ approaches zero, the quadratic region vanishes and approaches absolute error. Although lacking higher-order smoothness, this interpolation between classical losses makes Huber loss a widely used robust estimator. It serves as a stepping stone toward more aggressive redescending losses that suppress outliers more strongly, as discussed next.

3-4-2 Welsch Loss

Building on the bounded-influence Huber loss, the Welsch loss [32] (also called Leclerc loss) introduces a redescending robust penalty that suppresses large residuals through exponential downweighting. Unlike the Huber loss, it reduces the influence of large deviations without hard thresholds, making it suitable for scenarios with severe outliers.

For a scalar residual r, the Welsch loss is defined as

$$L_{\text{Welsch}}(r;c) = \frac{c^2}{2}\left(1 - \exp\left(-\frac{r^2}{c^2}\right)\right), \tag{3-37}$$

where $c > 0$ is a scaling parameter that controls the sensitivity to the residual magnitude. For small $|r| \ll c$, the function approximates a quadratic form, $L_{\text{Welsch}}(r) \approx \frac{r^2}{2}$. For large residuals, the loss saturates to $\frac{c^2}{2}$, bounding the maximum penalty and enhancing robustness.

The corresponding influence function, which dictates the sensitivity of the estimator to deviations, is obtained by differentiating Eq. (3-37) as follows:

$$\psi_{\text{Welsch}}(r;c) = \frac{\mathrm{d}}{\mathrm{d}r} L_{\text{Welsch}}(r;c) = r \exp\left(-\frac{r^2}{c^2}\right). \tag{3-38}$$

This function grows near the origin, attains a maximum, and then diminishes rapidly to zero as $|r| \to \infty$. This redescending behavior ensures that extreme residuals have almost no influence on the estimate, offering stronger suppression than Huber's bounded influence for extreme outliers.

Another advantage of the Welsch loss is its smoothness, as it is infinitely differentiable (C^∞) because of its exponential form, which makes it compatible with gradient-based algorithms that benefit from higher-order continuity. However, this loss is not globally convex. Its second derivative becomes negative beyond $|r| = c/\sqrt{2}$, introducing local concavity. Still, the scalar function $L_{\text{Welsch}}(r; c)$ remains unimodal, with a unique global minimum at $r = 0$ and no other stationary points.

In parameter estimation problems involving multiple residuals of the form $r_i(x) = z_i - H_i x$, the resulting objective $F(x) = \sum_i L(r_i(x))$ may become non-convex in the parameter space. This occurs when some residuals fall in the convex region (near zero), while others lie in the concave part, causing the overall Hessian to become indefinite. As a result, standard optimization methods may be stuck at saddle points or converge to suboptimal local minima. In practice, robust solvers require good initialization or continuation strategies similar to graduated nonconvexity to avoid poor convergence.

The Welsch loss provides strong robustness by combining bounded influence with smooth saturation, making it particularly effective in settings with extreme outlier contaminations. However, its nonconvexity presents challenges in frameworks that require provable convergence or tractable optimization. Despite its limitations, the Welsch loss serves as a natural precursor to even stricter outlier rejection strategies, such as the hard cutoff employed in Tukey's biweight loss.

3-4-3 Tukey's Biweight Loss

Similar to the Welsch loss, Tukey's biweight loss (also known as the bisquare loss) [26] is a redescending robust estimator. However, it enforces a stricter form of outlier suppression by completely rejecting residuals that exceed a predefined cutoff. In contrast to Welsch's gradual exponential decay, it results in zero influence beyond the threshold. It provides strong resistance to outliers, where extreme values are entirely excluded from influencing the estimate.

For a scalar residual r, the Tukey loss is defined as

$$L_{\text{Tukey}}(r; c) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{r}{c}\right)^2\right)^3\right], & |r| \le c, \\ \frac{c^2}{6}, & |r| > c, \end{cases} \tag{3-39}$$

where $c > 0$ is a constant that defines the rejection threshold. For residuals within $[-c, c]$, the function behaves like a polynomial, approximating a quadratic shape near the origin. For residuals larger than c, the loss saturates at a constant value $\frac{c^2}{6}$, thereby assigning no further penalty to increasingly extreme values.

The corresponding influence function, obtained by differentiating Eq. (3-39), is given by

$$\psi_{\text{Tukey}}(r;c) = \begin{cases} r\left(1-\left(\frac{r}{c}\right)^2\right)^2, & |r| < c, \\ 0, & |r| \geq c, \end{cases} \tag{3-40}$$

which smoothly tapers to zero at $|r| = c$ and remains zero for all larger residuals. This redescending behavior effectively eliminates the influence of all the gross outliers.

Although Tukey's loss is continuous and twice differentiable (C^2), it is not three times differentiable (C^3). The third derivative is discontinuous at the rejection boundary ($|r| = c$), and the function is non-convex because of its concave shoulder and flat plateau beyond the threshold. In the flat region, the gradient becomes zero, leading to a series of stationary points that incur the same costs. Although the global minimum of the scalar loss is still attained uniquely at $r = 0$, as with the Welsch loss, in parameter estimation problems with many residuals, the composite objective can be non-convex, potentially leading to saddle points or local minima depending on the data configuration.

This makes Tukey's biweight loss a powerful tool for heavily contaminated environments, achieving complete suppression of extreme outliers through a finite rejection threshold [33]. Nevertheless, its lack of convexity and restricted higher-order smoothness make optimization challenging and limit the application of theoretical tools that depend on global curvature.

3-4-4 Least Logarithmic Absolute Difference Loss

The LLAD loss [27] is a robust cost function introduced in the context of adaptive filtering to mitigate the sensitivity of traditional L_2- and L_1-based estimators to the outliers. Within a generalized logarithmic cost framework, the LLAD loss penalizes residuals logarithmically, combining the robustness of absolute loss with the smoothness of differentiable objectives.

For a scalar residual r, the LLAD loss is defined as

$$L_{\text{LLAD}}(r;\lambda) = \frac{1}{\lambda}\ln(1+\lambda|r|), \tag{3-41}$$

where $\lambda > 0$ is a tuning parameter that adjusts the transition between linear and sublinear behaviors. For small residuals $|r| \ll 1/\lambda$, the logarithm can be expanded via a Taylor series to yield $L_{\text{LLAD}}(r;\lambda) \approx |r| - \frac{\lambda}{2}|r|^2 + \mathcal{O}(|r|^3)$, indicating that the LLAD loss approximates an L_1 with a small quadratic correction near the origin. For large residuals $|r| \gg 1/\lambda$, the loss grows sub-linearly as

$$L_{\text{LLAD}}(r;\lambda) \approx \frac{1}{\lambda}\ln(\lambda|r|), \tag{3-42}$$

demonstrating a saturating behavior that bounds the penalty for extreme deviations. This diminishing sensitivity ensures that gross outliers contribute only marginally to the overall cost, thereby protecting the estimator from unbounded influence.

Differentiating Eq. (3-41) with respect to r yields the influence function

$$\psi_{\text{LLAD}}(r;\lambda) = \frac{\mathrm{d}}{\mathrm{d}r}L_{\text{LLAD}}(r;\lambda) = \frac{\text{sign}(r)}{1+\lambda|r|}, \tag{3-43}$$

which saturates to zero as $|r| \to \infty$. This redescending influence function limits the maximum impact of any single observation on the update, similar to the Welsch and Tukey losses.

Similar to other robust loss functions that exhibit redescending behavior, the LLAD loss is not globally convex. This is confirmed by the second derivative:

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2} L_{\mathrm{LLAD}}(r;\lambda) = -\frac{\lambda}{(1+\lambda|r|)^2}, \tag{3-44}$$

which is negative for all $r \neq 0$, indicating local concavity. Despite the loss being non-convex, it is unimodal with a single global minimum at $r = 0$. Furthermore, the absolute value introduces a non-differentiable point at $r = 0$. Although subgradients can be employed to address this issue in practice, their lack of smoothness limits the use of second-order methods.

While LLAD offers an elegant trade-off between efficiency and robustness, smoothly interpolating between the linear and logarithmic regimes to suppress the influence of large residuals, its non-convex nature and lack of smoothness at the origin restrict its application in distributed estimation frameworks.

3-4-5 Correntropy-Based Robustness and Kernel Risk-Sensitive Loss (KRSL)

Information-theoretic learning offers an alternative framework for robust estimation by redefining similarity through kernel-based measures [34]. It extends the redescending paradigms established by Welsch, Tukey, and LLAD losses, but instead of directly penalizing residual magnitudes, this approach quantifies the local similarity between the residual and zero using bounded kernel functions. This results in saturating loss profiles that naturally suppress outliers while preserving sensitivity near the origin.

A foundational concept in this framework is the MCC [29], which replaces conventional loss minimization with the maximization of the expected similarity, typically measured using a Gaussian kernel,

$$K_\sigma(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{3-45}$$

where $\sigma > 0$ is the kernel bandwidth parameter that controls the kernel width. This parameter defines the rate at which the kernel function decreases as r increases. Maximizing the expected kernel value is equivalent to minimizing the correntropy-induced loss,

$$L_{\mathrm{MCC}}(r;\sigma) = \sigma^2\left[1 - \exp\left(-\frac{r^2}{\sigma^2}\right)\right]. \tag{3-46}$$

This loss is mathematically equivalent, up to an additive constant and scaling, to the Welsch loss Eq. (3-37). Therefore, the MCC loss [28] inherits the same properties: it is infinitely differentiable (C^∞), redescending, unimodal at $r = 0$, and non-convex beyond $|r| > \sigma/\sqrt{2}$.

The KRSL was proposed in [30] as an extension of MCC that retains robustness while improving optimization properties. It reduces some non-convexity issues of MCC around the origin. The KRSL loss wraps the kernel discrepancy inside a second exponential function, yielding

$$L_{\mathrm{KRSL}}(r;\sigma,\lambda) = \frac{1}{\lambda}\exp\left[\lambda\left(1 - \exp\left(-\frac{r^2}{2\sigma^2}\right)\right)\right], \tag{3-47}$$

where $\lambda > 0$ is a risk-sensitive scalar that controls the sharpness of the outer exponential. As $\lambda \to 0$, the loss diverges like $1/\lambda$. For small residuals $|r| \ll \sigma$, Taylor expansion gives $L_{\text{KRSL}} \approx \frac{1}{\lambda} + \frac{r^2}{2\sigma^2} + \mathcal{O}(r^4)$, i.e., a constant offset plus a quadratic term. For large residuals $|r| \gg \sigma$, the inner kernel tends to zero, so the loss saturates at e^{λ}/λ, providing a limited penalty and thus strong outlier rejection.

The influence function for KRSL is given by

$$\psi_{\text{KRSL}}(r;\sigma,\lambda) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)\exp\left(\lambda\left[1-\exp\left(-\frac{r^2}{2\sigma^2}\right)\right]\right), \tag{3-48}$$

which remains bounded and redescending, but achieves a higher peak around moderate residuals compared to MCC. This enhances convergence by providing stronger gradients when residuals are within the inlier-to-outlier transition zone, without compromising robustness for large deviations. The KRSL loss, due to its exponential nature, is C^{∞} and shares non-convex characteristics as MCC, which can lead to optimization challenges. However, it is wider around the minimum and smoother, making convergence more reliable with proper initialization.

The selection of parameters σ and λ determines the balance between robustness and convergence. Increasing the parameter λ can speed up convergence, but also makes the cost landscape steeper. This can result in numerical instability, as large gradients might lead to overshooting or oscillations during optimization, particularly when using fixed or large step sizes. Computationally, KRSL adds one additional exponential evaluation per residual compared to MCC, which is negligible in small-scale problems but may increase runtime in large-scale or online filtering tasks.

The MCC loss functions like a Welsch-type redescending loss within an information-theoretic framework, offering bounded penalties resistant to outliers. The KRSL loss generalizes MCC by adding a tunable risk parameter that steepens the cost surface around moderate residuals, enhancing optimization while maintaining robustness. However, both losses remain non-convex and sensitive to hyperparameters, and their use in distributed or recursive settings requires careful consideration of computational costs and convergence.

3-5 Comparative Analysis and Rationale for Using Log-Cosh

In the proposed distributed robust estimation framework, the log-cosh loss function was selected after evaluating other robust loss functions. The aim is to achieve high resilience to outliers while maintaining feasible stability analysis and efficient distributed optimization. This requires that loss functions meet four criteria: (i) global convexity to ensure convergence to a single optimum; (ii) C^2 smoothness for conducting Hessian-based posterior analysis; (iii) a bounded influence to mitigate the impact of large residuals; and (iv) compatibility with distributed optimization methods, especially those using smooth gradient-based updates. Among robust losses examined, only the log-cosh loss fulfills all requirements, making it ideal for the framework.

Quadratic loss, also referred to as Mean Squared Error (MSE) or L_2 norm [35], is optimal under Gaussian noise and enjoys convexity and infinite differentiability. However, its unbounded influence function $\psi(r) = 2r$ causes extreme sensitivity to outliers. Large residuals

disproportionately dominate the objective and corrupt the estimates. This renders MSE infeasible in heavy-tailed noise environments, such as those encountered in WSNs with faulty or adversarial nodes.

The Mean Absolute Error (MAE), or L_1 norm [35], reduces sensitivity of quadratic loss by applying a linear penalty to residuals, bounding outlier influence. Although robust, the MAE loss is not optimal for nominal residuals. The Huber loss combines MSE sensitivity near origin with MAE robustness for large residuals by transitioning from quadratic to linear penalty at a fixed threshold. Both losses are convex; however, MAE is not differentiable at origin, and Huber loss, while C^1, lacks C^2 continuity at its transition point. The C^1 discontinuity makes MAE incompatible with gradient-based optimization algorithms, and the C^2 discontinuity rules out Laplace-based posterior approximations for both losses. As a result, uncertainty quantification is not applicable. The abrupt curvature change can destabilize step-size selection in gradient-based methods when iterates cross the threshold.

The LLAD loss applies a logarithmic transformation to the absolute residual, creating a saturating penalty that interpolates between L_1 and sublinear behavior. While the LLAD is continuous and differentiable in most regions, it is non-convex and lacks differentiability at the origin. Highly robust redescending losses, such as the Welsch and Tukey biweight, further mitigate outliers by assigning zero or near-zero influence to large residuals. These losses are also non-convex and feature flat plateaus in their objective surfaces. In a distributed setting, this introduces the risk of convergence to inconsistent local optima across nodes, thereby undermining consensus. These drawbacks limit the usefulness of the LLAD in rigorous distributed inference.

Information-theoretic losses, such as the MCC and KRSL, introduce kernel-based similarity measures that perform well in non-Gaussian or impulsive noise settings. MCC employs a Gaussian kernel to bound the cost of large errors, whereas KRSL wraps the correntropy term in an additional exponential to steepen the surface near the origin and accelerate empirical convergence. Despite their C^∞ smoothness, both losses remain non-convex and sensitive to the kernel-width parameter, which must align with the noise scale. The objective landscapes can contain multiple local minima and flat regions, making performance depend on careful initialization and step size.

In contrast, the log-cosh loss combines key theoretical and practical advantages. It is strictly convex, globally smooth (C^∞), and has a bounded influence function $\psi(r) = \tanh(\alpha r)$. This structure ensures nodes converge to a unique global optimum using gradient updates. The second derivative, $\psi'(r) = \alpha\,\mathrm{sech}^2(\alpha r)$, is strictly positive and smoothly decaying, forming a well-conditioned and positive-definite Hessian that supports Laplace-based covariance approximation and stability analysis. This curvature enables the robust filter to be modeled as a KF with intermittent observations, allowing for proving bounded estimation error. The log-cosh loss integrates into the EXTRA consensus framework. Its smooth, convex objective allows nodes to compute local gradients and reach agreement using neighbor communication.

To conclude, the log-cosh loss is uniquely equipped to meet all essential criteria for robust distributed estimation in WSNs. It effectively mitigates the impact of outliers, guarantees unique minima, facilitates analytical stability proofs and posterior covariance approximation, and can be practically implemented to achieve accurate consensus. No traditional or information-theoretic alternatives meet this comprehensive set of criteria. Therefore, the

adoption of the log-cosh loss is not simply incidental; it is crucial to the theoretical and computational developments discussed in this work.

3-6 Distinction From Prior Work

Although the log-cosh loss has been investigated in the context of robust signal processing, its application has been limited to centralized adaptive filtering techniques. Most notable examples include the Least log-cosh Algorithm (LLA) [36] and the Logarithmic Hyperbolic Cosine Adaptive Filter (LHCAF) [37], both designed to improve resilience in single-sensor environments affected by impulsive noise. These approaches show the effectiveness of the log-cosh function in effectively mitigating large residuals due to its non-quadratic growth characteristics. However, these prior studies fundamentally differ from the proposed framework in terms of the problem domain, theoretical examination, and implementation architecture.

The LLA and LHCAF algorithms are designed for static system identification, aiming to estimate a constant set of model parameters like filter coefficients. Their evaluation centers on metrics such as Mean Square Deviation (MSD) and Excess Mean Square Error (EMSE), with a focus on achieving convergence under stationary conditions. In contrast, this thesis addresses the dynamic challenge of distributed state estimation, where the latent state changes over time and must be estimated recursively from noisy local observations.

To ensure robustness with outliers, a generalized Bayesian inference framework is adopted, substituting conventional Gaussian log-likelihood with log-cosh loss. This approach integrates robustness into the posterior distribution in the probabilistic framework, aiding in deriving update equations and facilitating systematic posterior uncertainty quantification using Laplace-based covariance approximations. While generalized Bayesian methods have been examined in robust inference literature [38, 15], the proposed framework extends this concept to dynamic and distributed state estimation, presenting challenges including recursive uncertainty propagation, distributed systems implementation, and maintaining bounded error covariance over time.

Unlike prior centralized methods, the framework proposed in this thesis is designed for distributed implementation in WSNs, where nodes compute local updates and communicate with neighbors. The implementation uses the EXTRA algorithm, enabling exact convergence in convex consensus problems using local gradients and neighbour communication. To provide analytical guarantees, stability analysis is performed by modeling the robust update as a KF with intermittent observations [39]. This approach applies known results from switching system theory to derive conditions for bounded estimation error. Such guarantees are critical in distributed estimation, where local instability can degrade global performance.

While both the log-cosh loss and generalized Bayes have been examined separately, their integration into a cohesive, scalable, and demonstrably robust filtering framework represents a novel approach. This new method combines the design of robust loss, consensus optimization through EXTRA, and stability analysis to offer a comprehensive solution for distributed estimation in the presence of adversarial noise. This integration addresses challenges that have not been previously addressed in earlier uses of the log-cosh loss and constitutes the main contribution of this thesis. The entire work is detailed as a paper in the following section.

Robust Estimation in Fully Distributed Sensor Network in the Presence of Outliers

Hardik Aggarwal, Chen Quan

***Abstract*—Classical distributed estimation algorithms for state estimation in Wireless Sensor Networks (WSNs), such as consensus-based Kalman filtering and diffusion strategies, typically assume Gaussian observation models, under which outliers are rare. However, even a few such outliers can significantly skew local updates and degrade the accuracy of the global estimate. Moreover, these networks may suffer from unreliable measurements and limited communication budgets, further challenging the robustness and effectiveness of traditional estimation methods. To overcome these limitations, we propose a novel estimation framework that leverages generalized Bayesian inference to design robust loss functions, enabling sensor nodes to perform outlier-resilient local updates. Specifically, the log-cosh loss is employed as a smooth, convex alternative to the log-likelihood-based squared penalty on innovation residuals, which limits the influence of extreme values while preserving differentiability. The resulting robust update is embedded within a recursive filtering structure and implemented using the Exact First Order Algorithm (EXTRA) consensus optimization algorithm. Moreover, we conduct a stability analysis demonstrating that it maintains a bounded error covariance provided the designed robustness parameter satisfies an analytically derived condition. To enhance performance without sacrificing stability, an adaptive strategy is introduced that dynamically adjusts the robustness parameter based on the residual magnitude. Theoretical analysis and numerical results demonstrate that the proposed approach achieves accurate and resilient state estimation in resource-constrained and adversarial environments, yielding 5-10% lower average RMSE in the presence of measurement outliers and up to 20% improvement when additional process disturbances are introduced, while also requiring fewer consensus iterations for convergence.**

I. Introduction

Reliable state estimation is essential for enabling intelligent decision-making, anomaly detection, and control across a broad spectrum of applications, including environmental monitoring, industrial automation, and structural health assessment [1], [2]. This task involves inferring latent global parameters from spatially distributed, noise-corrupted measurements collected by sensors deployed throughout the network.

Early estimation techniques for WSNs primarily relied on centralized architectures, where all the local measurements were aggregated at a Fusion Center (FC) to compute a global estimate of the parameter of interest. Several classic estimation techniques have been proposed for such networks, including the Centralized Kalman Filter (KF) [3] and Centralized Least Mean Squares (CLMS) [4]. Although this setup provides access to the complete dataset and facilitates globally optimal solutions under ideal conditions [5], it imposes significant practical limitations. The FC constitutes a single point of failure, and communication overhead scales poorly with network size, leading to rapid energy depletion in power-constrained nodes [6].

To alleviate these challenges, decentralized wireless sensor networks emerged as a scalable alternative that enables real-time estimation without centralized data fusion. In decentralized networks, nodes are organized hierarchically, and designated cluster heads or intermediate aggregators perform local data fusion. This approach improves scalability and balances the communication load during the estimation process. Classical estimation techniques for such networks have been proposed in [7], [8]. However, this decentralized architecture remains susceptible to structural fragilities. Failures at intermediate nodes can disrupt estimation for entire subnetworks, and multi-hop communication introduces additional latency and coordination complexity [9].

To overcome the aforementioned limitations, fully distributed estimation has emerged as a promising approach that eliminates the need for centralized and hierarchical coordination. In fully distributed networks, each node performs local computations and communicates only with its immediate neighbors. Through iterative communication within these neighborhoods, nodes collaboratively refine their estimates toward a consistent approximation of the global state. Algorithms based on consensus [10], [11], diffusion [12], and gossip [13] exemplify this paradigm. Moreover, distributed estimation offers substantial benefits in energy efficiency, scalability, and robustness to node failures, rendering it well-suited for modern WSNs deployments characterized by dynamic, large-scale, and resource-constrained environments.

Building upon foundational consensus and diffusion protocols, various distributed estimation algorithms have been developed to address challenges such as optimality, partial observability, and unknown cross-covariance of estimation errors arising from inter-node information exchange. For instance, the Distributed Kalman Filter (DKF) is proposed in [11], [14] for distributed networks as an extension of the centralized KF [3] by enabling each node to perform local updates and fuse estimates with neighbors through consensus [15]. To handle sparse sensing, the Generalized Kalman Consensus Filter (GKCF) proposed in [16] incorporates inverse covariance weights in consensus, improving performance at naive nodes, i.e., nodes that lack direct observations of the target. When cross-covariances are unknown, conservative strategies such as Covariance Intersection (CI) proposed in [17] provided consis-

Hardik Aggarwal is with the Delft Center for Systems and Control, TU Delft, 2628 CD Delft, Netherlands. Emails: h.aggarwal@student.tudelft.nl.

C. Quan is with the Faculty of Electrical Engineering, Mathematics, and Computer Science, TU Delft, 2628 CD Delft, Netherlands. Emails: c.quan@tudelft.nl.

tent fusion without assuming independence. The Hybrid Consensus on Measurement and Information (HCMCI) proposed in [18] combined CI-based fusion of priors with consensus-based update integration. A Diffusion technique proposed in [19] reduced communication by exchanging compressed summaries, preserving convergence with lower resource usage. A Gossip-based methods proposed in [20] relies on randomized pairwise exchanges and operate asynchronously, offering robustness to delays and topology changes.

Despite the advantages of distributed estimation algorithms based on consensus, diffusion, and gossip protocols, ensuring robustness against outliers remains a significant challenge [21]. Outliers, arising from sensor faults, environmental disturbances, or adversarial attacks, can severely degrade performance, especially in distributed settings [22]. As information is iteratively exchanged among neighboring nodes, corrupted measurements can contaminate local estimates and propagate throughout the network, leading to biased or unstable global behavior. Thus, robust estimation mechanisms that can effectively mitigate the impact of anomalous data are needed in distributed settings. However, the aforementioned algorithms [10]-[20], which rely on Gaussian noise assumptions, assign negligible probability to extreme deviations, making them unsuitable for heavy-tailed or impulsive measurement noise [23], and thus lacking effective mechanisms to suppress anomalies.

To overcome this limitation, recent efforts have focused on developing robust distributed estimation techniques that tolerate outliers and account for model mismatch, where the assumed noise distribution does not reflect the actual measurement statistics. For instance, the Robust Consensus Nonlinear Information Filter (RCNIF) proposed in [24] addressed measurement corruption by modeling observation noise using a heavy-tailed Student-t distribution and employing a Variational Bayesian (VB) framework to jointly infer the state and latent noise parameters. Similarly, the Diffusion Minimum Generalized Rank Norm (dMGRN) proposed in [25] incorporated rank-based statistics through the Generalized Rank (GR) norm and robust weighting using the Minimum Volume Ellipsoid (MVE), improving resilience to outliers in both input and output spaces. While effective, these methods typically involve iterative variational inference or non-convex optimization, limiting scalability in real-time or large-scale networks. Alternatively, in [26] the Weighted Observation Likelihood Filter (WoLF) algorithm was proposed that introduced robustness via generalized Bayesian inference using a weighted loss function in place of the log-likelihood. This formulation retained closed-form updates and improved efficiency, but its centralized design restricted its applicability in decentralized sensor networks.

In this context, robust loss functions have emerged as an effective approach to mitigating the influence of outliers in estimation. Unlike the classical Mean Squared Error (MSE), which is optimal under Gaussian noise but disproportionately penalizes large residuals, robust alternatives grow subquadratically with the error magnitude, limiting the influence of extreme observations. Common examples include the L1 norm used in Mean Absolute Error (MAE) [27], the Huber loss [28] that transitions from quadratic to linear growth beyond a threshold δ, and the Least Logarithmic Absolute Difference (LLAD) [29], which applies a saturating logarithmic penalty. These losses offer distinct trade-offs between robustness, convergence behavior, and ease of optimization. For instance, MAE is highly robust but non-differentiable at zero; Huber loss remains mostly convex and smooth but requires careful tuning; LLAD enhances performance under impulsive noise but introduces non-convexity.

Another class of robust losses is based on information-theoretic principles. Correntropy, which defines a kernel-based similarity between the prediction error and zero, effectively suppresses outliers by assigning a negligible cost to large deviations beyond the kernel bandwidth [30], [31]. However, its loss surface is typically non-convex, with steep curvature near zero and flat tails, which can hinder convergence in gradient-based algorithms. To address this, the Kernel Risk-Sensitive Loss (KRSL) proposed in [32] incorporates a risk-sensitive criterion into a reproducing kernel Hilbert space, yielding a smoother and more convex objective. Although KRSL introduces additional tuning parameters and higher computational cost, it improves convergence speed and estimation accuracy over standard correntropy approaches. These robust loss formulations provide a foundation for developing distributed estimators that remain resilient under non-Gaussian and adversarial noise.

In this paper, we propose a robust estimation framework for distributed state estimation in WSNs that integrates the log-cosh loss function into a generalized Bayesian inference formulation. This choice is key to enabling both provable stability and efficient fully distributed implementation. While several robust loss functions have been studied, including the L1 norm, Huber loss, LLAD, and KRSL, each exhibits specific limitations that hinder either distributed implementation or theoretical analysis. The L1 norm and Huber loss are not twice differentiable, complicating posterior approximation and stability proofs. LLAD and correntropy are non-convex, which limits convergence guarantees and makes the analysis of stability intractable. KRSL also introduce kernel parameters that affect performance and add computational complexity. These structural and analytical limitations prevent direct application in dynamic distributed settings.

The log-cosh loss combines three properties that are essential to our framework: convexity, infinite differentiability, and a globally Lipschitz continuous gradient. These characteristics enable tractable second-order posterior approximation, a rigorous stability analysis by modeling the estimator as a KF with intermittent observations, and an efficient implementation using the EXTRA algorithm without inner-loop optimization or dual variables. While the hyperbolic cosine function has previously been employed in robust adaptive filtering, most notably in the least log-cosh algorithm [33] and the Logarithmic Hyperbolic Cosine Adaptive Filter (LHCAF) [34], these methods address a fundamentally different class of problems. Specifically, they focus on static system identification in centralized architectures and study weight convergence, rather than dynamic state estimation or stability. Moreover, they do not support distributed implementation or provide formal

guarantees on estimation error. By contrast, our approach is tailored to state estimation problems based on distributed sensor measurements, with formal guarantees on estimation stability. The main contributions of this work are as follows:

1) We formulate a robust recursive estimation algorithm by integrating the log-cosh loss into a generalized Bayesian framework. This yields a Kalman filter-like recursion that suppresses outliers while preserving differentiability and convexity for efficient optimization.
2) We establish stability analysis by conservatively modeling the robust filter as a Kalman filter with intermittent observations, providing theoretical guarantees on bounded estimation error when the robustness parameter satisfies an certain condition.
3) We derive an explicit design criterion for the robustness parameter that can balance robustness against nominal performance, while still ensuring the stability of the system.
4) We implement the proposed robust filter in a fully distributed setting using the EXTRA for consensus optimization. Convergence is achieved using only local gradients and fixed step sizes, without relying on inner-loop optimization or dual variables, thereby reducing computational demand.

Throughout this work, vectors and matrices are represented by bold lowercase (e.g., $\mathbf{x}_k$) and bold uppercase (e.g., $\mathbf{A}_k$) letters, respectively. Bold calligraphic symbols (e.g., $\boldsymbol{\mathcal{Z}}_k$, $\boldsymbol{\mathcal{H}}_k$) denote globally aggregated quantities obtained by stacking local contributions. The predicted and updated estimates at time k are denoted by $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$, respectively. The value of a variable at iteration ℓ of a consensus process is denoted by $\mathbf{x}^{(\ell)}$. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The infinity norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is denoted by $\|\mathbf{v}\|_\infty = \max_{i=1,\dots,n} |v_i|$.

The remainder of the paper is organized as follows. Section II introduces the system model and problem formulation for distributed state estimation in sensor networks. Section III presents the proposed robust local filtering framework based on generalized Bayesian updating with the log-cosh loss. Section IV describes the fully distributed implementation using the EXTRA algorithm for consensus optimization. In Section V, we establish the stability properties of the proposed robust filter and derive a design rule for the robustness parameter. Section VI reports numerical results demonstrating the efficacy and resilience of the proposed method. Finally, Section VII concludes the paper.

II. Problem Formulation

Consider a fully distributed WSN comprising N sensor nodes, whose goal is to estimate the dynamic state of a physical process. A representative application is the tracking of a mobile target, such as a ground vehicle or unmanned aerial vehicle (UAV), within a surveillance region. All sensor nodes, each with limited sensing and communication capabilities, collaboratively estimate the global state in the absence of a centralized FC. In the remainder of this section, we present the dynamical system model and the underlying network topology that characterize the neighborhood communication in the absence of the FC.

A. System Model

Let $\mathbf{x}_k \in \mathbb{R}^n$ denote the global state at discrete time step k, where n is the state dimension. In a typical two-dimensional tracking scenario, the state may consist of position and velocity components:

$$\mathbf{x}_k = \begin{bmatrix} x_k & y_k & \dot{x}_k & \dot{y}_k \end{bmatrix}^\top,$$

where (x_k, y_k) and $(\dot{x}_k, \dot{y}_k)$ represent the target's position and velocity, respectively. The state evolves according to the discrete-time linear model

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{w}_k, \tag{1}$$

where $\mathbf{A}_k \in \mathbb{R}^{n\times n}$ is the state transition matrix, and $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ is zero-mean Gaussian process noise with covariance $\mathbf{Q}_k \in \mathbb{R}^{n\times n}$.

Each node $i \in \mathcal{V}$ obtains a local measurement described by

$$\mathbf{z}_{i,k} = \mathbf{H}_{i,k} \mathbf{x}_k + \boldsymbol{\nu}_{i,k}, \tag{2}$$

where $\mathbf{z}_{i,k} \in \mathbb{R}^{m_i}$ and $\mathbf{H}_{i,k} \in \mathbb{R}^{m_i \times n}$ denote the observation vector and the local observation matrix of node i at time step k, respectively. The measurement noise $\boldsymbol{\nu}_{i,k}$ at node i and time step k is modeled as a zero-mean random vector with known, positive-definite covariance $\mathbf{R}_{i,k} \in \mathbb{R}^{m_i \times m_i}$. Although the true distribution of $\boldsymbol{\nu}_{i,k}$ may be non-Gaussian due to contamination by outliers, it is approximated as Gaussian with covariance $\mathbf{R}_{i,k}$ within the *inlier region*. This region is defined as the subset of the measurement space where observations are expected to lie in the absence of outliers and are consistent with typical system behavior. Within this region, the second-order statistics of $\boldsymbol{\nu}_{i,k}$ are well approximated by a Gaussian model with covariance $\mathbf{R}_{i,k}$. We assume that the process noise $\mathbf{w}_k$ and the measurement noises $\boldsymbol{\nu}_{i,k}$ are mutually independent across all nodes $i \in \mathcal{V}$ and time steps k. Furthermore, the system matrices $\mathbf{A}_k$, $\mathbf{H}_{i,k}$, $\mathbf{Q}_k$, and $\mathbf{R}_{i,k}$ are assumed to be known, may vary with time, and are deterministic.

B. Network Topology

The communication topology among the sensor nodes, modeled as vertices $\mathcal{V} = \{1, \dots, N\}$ in an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes bidirectional communication links. Each node $i \in \mathcal{V}$ communicates with its closed neighborhood

$$\mathcal{N}_i = \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\} \cup \{i\}, \tag{3}$$

which includes the node itself and all directly connected neighbors. The structure of the network can be algebraically described using the adjacency matrix $\mathbf{A}_{\text{adj}} \in \mathbb{R}^{N\times N}$, where $[\mathbf{A}_{\text{adj}}]_{ij} = 1$ if $(i,j) \in \mathcal{E}$, and zero otherwise. The degree matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is diagonal with entries $[\mathbf{D}]_{ii} = \sum_{j=1}^{N} [\mathbf{A}_{\text{adj}}]_{ij}$, indicating the number of neighbors of node i. The graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}_{\text{adj}}$ and plays a central role in analyzing the convergence of consensus algorithms. The matrix $\mathbf{L}$ is symmetric and positive semidefinite,

with eigenvalues ordered as $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. The smallest eigenvalue $\lambda_1 = 0$ corresponds to the eigenvector $\mathbf{1}_N$, the vector of all ones, and its multiplicity equals the number of connected components in the graph. Therefore, the graph $\mathcal{G}$ is connected if and only if the zero eigenvalue has multiplicity one, which implies $\lambda_2 > 0$, where λ_2 is referred to as the *algebraic connectivity* of the graph [35].

In the context of distributed estimation, the Laplacian governs the convergence dynamics of consensus protocols. Specifically, many consensus-based algorithms update local variables through weighted averaging governed by the Laplacian matrix. The rate at which consensus is achieved depends on the spectral properties of $\mathbf{L}$, particularly the value of λ_2. A larger λ_2 implies stronger graph connectivity and results in faster convergence to the consensus value.

III. Robust Estimation

In this section, we first review the centralized KF, which provides the Minimum Mean Squared Estimate (MMSE) for the global state under linear-Gaussian assumptions. We then introduce a robust local update framework based on generalized Bayesian inference with log-cosh loss, and develop a recursive filtering formulation that retains the structure of the Kalman filter while improving resilience to outliers. The resulting robust local filter serves as a foundational component in the proposed distributed estimation algorithm.

A. Centralized Kalman Filter

The centralized estimator recursively applies a two-stage prediction-update cycle based on the state and observation models defined in (1) and (2). The prediction step is given by

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}_k \hat{\mathbf{x}}_{k-1|k-1}, \tag{4}$$

$$\mathbf{P}_{k|k-1} = \mathbf{A}_k \mathbf{P}_{k-1|k-1} \mathbf{A}_k^\top + \mathbf{Q}_k, \tag{5}$$

and the update step is

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \boldsymbol{\mathcal{H}}_k^\top \left(\boldsymbol{\mathcal{H}}_k \mathbf{P}_{k|k-1} \boldsymbol{\mathcal{H}}_k^\top + \boldsymbol{\mathcal{R}}_k \right)^{-1}, \tag{6}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \left(\boldsymbol{\mathcal{Z}}_k - \boldsymbol{\mathcal{H}}_k \hat{\mathbf{x}}_{k|k-1} \right), \tag{7}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \boldsymbol{\mathcal{H}}_k) \mathbf{P}_{k|k-1}. \tag{8}$$

Here, $\boldsymbol{\mathcal{Z}}_k = \begin{bmatrix} \mathbf{z}_{1,k}^\top & \cdots & \mathbf{z}_{N,k}^\top \end{bmatrix}^\top \in \mathbb{R}^{m_{\text{tot}}}$ denotes the global measurement vector, where $m_{\text{tot}} = \sum_{i=1}^N m_i$ is the total measurement dimension across the network. The matrix $\boldsymbol{\mathcal{H}}_k = \begin{bmatrix} \mathbf{H}_{1,k}^\top & \cdots & \mathbf{H}_{N,k}^\top \end{bmatrix}^\top \in \mathbb{R}^{m_{\text{tot}} \times n}$ is the global observation matrix. The matrix $\boldsymbol{\mathcal{R}}_k \in \mathbb{R}^{m_{\text{tot}} \times m_{\text{tot}}}$ is block-diagonal, formed by aggregating the local noise covariances $\mathbf{R}_{i,k}$, reflecting the assumption that measurement noise is uncorrelated across nodes.

B. Robust Local Update via Generalized Bayes Rule

In the classical KF, each local measurement update is based on the assumption that the measurement noise is Gaussian. Under this model, the posterior distribution of the state $\mathbf{x}_k$ is given by Bayes rule as

$$p(\mathbf{x}_k \mid \mathbf{z}_{i,k}) \propto p(\mathbf{x}_k)\, p(\mathbf{z}_{i,k} \mid \mathbf{x}_k), \tag{9}$$

where $p(\mathbf{x}_k)$ is the predicted prior distribution obtained by propagating the previous state's posterior through the system dynamics model, and $p(\mathbf{z}_{i,k} \mid \mathbf{x}_k)$ is the likelihood of the local measurement $\mathbf{z}_{i,k}$ given the state $\mathbf{x}_k$. When the measurement noise $\boldsymbol{\nu}_{i,k}$ in (2) is modeled as zero-mean Gaussian with covariance $\mathbf{R}_{i,k}$, the likelihood $p(\mathbf{z}_{i,k} \mid \mathbf{x}_k)$ is given by

$$p(\mathbf{z}_{i,k} \mid \mathbf{x}_k) \propto \exp\Bigg\{ -\frac{1}{2} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right)^\top \mathbf{R}_{i,k}^{-1} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right) \Bigg\}.$$

The estimate of the state $\mathbf{x}_k$ can be obtained with the Maximum A Posteriori (MAP) estimator, which maximizes the posterior distribution $p(\mathbf{x}_k \mid \mathbf{z}_{i,k})$, or equivalently, minimizes the negative logarithm $-\log(p(\mathbf{x}_k \mid \mathbf{z}_{i,k}))$. The log-posterior $\log(p(\mathbf{x}_k \mid \mathbf{z}_{i,k}))$ consists of two components: the log-prior and the log-likelihood as expressed below:

$$\log(p(\mathbf{x}_k \mid \mathbf{z}_{i,k})) = \underbrace{\log(p(\mathbf{x}_k))}_{\text{log-prior}} + \underbrace{\log(p(\mathbf{z}_{i,k} \mid \mathbf{x}_k))}_{\text{log-likelihood}}. \tag{10}$$

The negative of the latter term, $-\log(p(\mathbf{z}_{i,k} \mid \mathbf{x}_k))$, also referred to as the negative log-likelihood, represents the loss function that penalizes the discrepancy between the predicted observation $\mathbf{H}_{i,k}\mathbf{x}_k$ and the actual measurement $\mathbf{z}_{i,k}$. Under a Gaussian noise model, this term reduces to the standard quadratic loss, up to an additive constant:

$$-\log(p(\mathbf{z}_{i,k} \mid \mathbf{x}_k)) = \frac{1}{2} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right)^\top \mathbf{R}_{i,k}^{-1} \left(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_k\right). \tag{11}$$

While statistically optimal under Gaussian noise, this formulation's reliance on a quadratic penalty causes unbounded growth for large residuals, making the estimator highly sensitive to outliers that can dominate the state update.

To mitigate this sensitivity, we adopt a robust statistical framework that preserves the Bayesian update structure while attenuating the effect of extreme residuals. The principled approach, as proposed in [36], is to generalize the update by directly replacing the negative log-likelihood with a robust loss function $L_{i,k}(\mathbf{x}_k)$. This function is selected for its properties, such as bounded influence, rather than being derived from a specific probabilistic measurement model. This leads to the generalized Bayes rule, where the generalized posterior is defined as

$$p(\mathbf{x}_k \mid \mathbf{z}_{i,k}) \propto p(\mathbf{x}_k)\, \exp\left\{-L_{i,k}(\mathbf{x}_k)\right\}. \tag{12}$$

This generalized formulation can enhance robustness against heavy-tailed noise and other anomalous measurements when an appropriate loss is chosen. In this work, we adopt the log-cosh loss, which behaves quadratically for small residuals (much like the standard Gaussian model) but smoothly transitions to a linear penalty for large deviations. This property mitigates the influence of extreme outliers on the accuracy of the state estimate.

Each sensor node $i \in \mathcal{V}$ obtains a local measurement $\mathbf{z}_{i,k} \in \mathbb{R}^{m_i}$ at time k according to the model (2). Under the

robust formulation, each node constructs a loss based on its measurement residual, defined as a function of the state:

$$\mathbf{r}_{i,k}(\mathbf{x}_k) = \mathbf{z}_{i,k} - \mathbf{H}_{i,k}\,\mathbf{x}_k. \tag{13}$$

To account for heteroscedastic measurement noise, the residual is whitened with respect to the measurement covariance using its inverse square root:

$$\tilde{\mathbf{r}}_{i,k}(\mathbf{x}_k) = \mathbf{R}_{i,k}^{-\frac{1}{2}}\,\mathbf{r}_{i,k}(\mathbf{x}_k), \tag{14}$$

where $\mathbf{R}_{i,k}$ is the measurement noise covariance under inlier conditions, as defined in Section II. The log-cosh loss is then applied independently to each component of the whitened residual vector $\tilde{\mathbf{r}}_{i,k}(\mathbf{x}_k)$, yielding the robust local loss

$$L_{i,k}(\mathbf{x}_k) = \frac{1}{\alpha^2}\sum_{l=1}^{m_i} \log\cosh\left(\alpha\,\tilde{r}_{i,k,l}(\mathbf{x}_k)\right), \tag{15}$$

where $\alpha > 0$ is a robustness tunable parameter and $\tilde{r}_{i,k,l}(\mathbf{x}_k)$ denotes the l-th component of the whitened residual vector. The parameter α appears inside the log-cosh function to control the trade-off between robustness and fidelity. Specifically, increasing α narrows the range over which the loss behaves quadratically, causing it to transition to linear growth at smaller residual magnitudes, which improves robustness to outliers. Decreasing α retains the quadratic approximation over a wider range, which improves statistical efficiency under nominal noise conditions. To formalize these qualitative observations, we now characterize the asymptotic behavior of the scaled log-cosh loss in the inlier and outlier regimes in Proposition 1.

Proposition 1. *The scaled log-cosh loss exhibits the following asymptotic behavior:*

$$\frac{1}{\alpha^2}\log\cosh(\alpha\tilde{r}) = \frac{1}{2}\tilde{r}^2 + \mathcal{O}(\tilde{r}^4) \quad \textit{as } \tilde{r} \to 0, \tag{16}$$

$$\frac{1}{\alpha^2}\log\cosh(\alpha\tilde{r}) = \frac{1}{\alpha}|\tilde{r}| - \frac{\log 2}{\alpha^2} + \mathcal{O}(e^{-2\alpha|\tilde{r}|}) \quad \textit{as } |\tilde{r}| \to \infty. \tag{17}$$

where, $\tilde{r} \in \mathbb{R}$ denotes a component of the whitened residual vector $\tilde{\mathbf{r}}_{i,k}(\mathbf{x}_k)$.

Proof: For small arguments, the Taylor expansion of $\log\cosh(x)$ around $x = 0$ yields

$$\log\cosh(x) = \frac{x^2}{2} - \frac{x^4}{12} + \mathcal{O}(x^6).$$

Dividing by α^2 and substituting $x = \alpha\tilde{r}$ gives

$$\frac{1}{\alpha^2}\log\cosh(\alpha\tilde{r}) = \frac{1}{2}\tilde{r}^2 - \frac{\alpha^2\tilde{r}^4}{12} + \mathcal{O}(\alpha^4\tilde{r}^6),$$

which confirms the approximation in (16).

For large arguments, we use the identity

$$\log\cosh(x) = |x| - \log 2 + \mathcal{O}(e^{-2|x|}) \quad \text{as } |x| \to \infty.$$

Substituting $x = \alpha\tilde{r}$ and dividing by α^2 yields

$$\frac{1}{\alpha^2}\log\cosh(\alpha\tilde{r}) = \frac{1}{\alpha}|\tilde{r}| - \frac{\log 2}{\alpha^2} + \mathcal{O}(e^{-2\alpha|\tilde{r}|}),$$

establishing the asymptotic behavior in (17). ■

The small-residual approximation in (16) confirms that, in the inlier regime, the loss behaves nearly identically to the standard quadratic form. The scaling factor $\frac{1}{\alpha^2}$ is chosen to cancel the leading-order effect of α, yielding the approximation $\frac{1}{2}\tilde{r}^2$, and thereby ensuring consistency with classical Kalman filtering under nominal Gaussian noise. Conversely, the large-residual approximation in (17) shows that the loss transitions smoothly to a linear growth regime. This bounded influence property limits the impact of extreme measurements, enhancing robustness under heavy-tailed noise and sensor faults.

Beyond this asymptotic behavior, the log-cosh loss exhibits several structural features that make it particularly well-suited for efficient optimization within robust estimation frameworks. Defined as $f(x) = \log\cosh(x)$, the loss is strictly convex and infinitely differentiable on $\mathbb{R}$, with a globally Lipschitz continuous gradient. Its first derivative, $f'(x) = \tanh(x)$, is bounded in magnitude by one, which automatically limits the influence of large residuals by saturating the gradient. This saturation prevents overly aggressive steps and supports the use of relatively large learning rates in gradient-based methods without risking instability. Additionally, the second derivative $f''(x) = \mathrm{sech}^2(x)$ is strictly positive and smooth, enabling the application of Newton and quasi-Newton optimization techniques to accelerate convergence. In contrast to non-smooth alternatives such as the L_1 norm or the Huber loss, the log-cosh loss maintains continuous curvature, simplifying algorithmic implementation and improving convergence behavior.

Although the curvature $f''(x)$ vanishes asymptotically, indicating that $f(x)$ is not strongly convex across its entire domain, the combined objective in (19) remains strongly convex due to the contribution of the quadratic prior. This favorable structure strikes a balance between statistical robustness and numerical efficiency. For further discussion of these properties, see [37], [38].

The resulting local loss provides a robust and consistent foundation for the subsequent fusion and consensus operations in the distributed estimation process (see Section IV).

C. Recursive Robust Estimation via Local Filtering

We now formulate a recursive estimation procedure that incorporates the robust local update derived in the preceding subsection. This procedure defines a robust filter operating at an individual sensor node, without relying on any network-level communication or data fusion. The objective is to isolate the effect of the robust measurement update on the local recursive structure, thereby facilitating a clearer understanding of its influence on estimation performance in the presence of outliers. The resulting local filter forms a foundational component for subsequent integration into a fully distributed estimation algorithm. For notational simplicity in this section, we will temporarily omit the agent index i when defining the local filter structure.

The proposed robust filter for the single-node scenario follows the standard two-stage recursion: prediction and update. In the prediction stage, the prior estimate of the state $\hat{\mathbf{x}}_{k|k-1}$

and its associated error covariance $\mathbf{P}_{k|k-1}$ at time step k are computed using the prediction equations (4) and (5). In the update stage, the filter incorporates the new measurement $\mathbf{z}_k \in \mathbb{R}^m$, where $\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k$, to refine the state estimate.

To formalize the update step, we invoke the generalized Bayes rule in (12), where the posterior is proportional to the product of a Gaussian prior and the exponential of a robust loss function. Taking the negative logarithm and omitting additive terms that are independent of $\mathbf{x}_k$, the posterior state estimate is given by the solution to the following minimization problem:

$$\hat{\mathbf{x}}_{k|k} = \arg\min_{\mathbf{x}_k \in \mathbb{R}^n} J_k(\mathbf{x}_k), \tag{18}$$

where the cost function is defined as

$$\begin{aligned} J_k(\mathbf{x}_k) &= -\log p(\mathbf{x}_k) + L_k(\mathbf{x}_k) \\ &= \frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^\top \mathbf{P}_{k|k-1}^{-1}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) \\ &\quad + \frac{1}{\alpha^2}\sum_{l=1}^{m} \log\cosh\left(\alpha\, \tilde{r}_{k,l}(\mathbf{x}_k)\right), \end{aligned} \tag{19}$$

with $\tilde{r}_{k,l}(\mathbf{x}_k)$ defined in (14). The first term in (19) encodes the Gaussian prior from the prediction step, while the second term introduces robustness by penalizing large residuals sub-quadratically, thereby mitigating the influence of outliers. To ensure the well-posedness of the optimization problem (18), we first establish the existence and uniqueness of its solution.

Lemma 1 (Existence and Uniqueness of the MAP Estimate)**.** *The cost function $J_k(\mathbf{x}_k)$ defined in* (19) *is strongly convex and coercive. Consequently, it admits a unique minimizer, which we define as the robust MAP estimate $\hat{\mathbf{x}}_{k|k}$.*

Proof: The cost function $J_k(\mathbf{x}_k)$ is the sum of two terms. The first is a strictly convex quadratic form weighted by the positive definite matrix $\mathbf{P}_{k|k-1}^{-1} \succ 0$, and is therefore strongly convex. The second term is a sum of convex log-cosh functions applied to affine transformations of $\mathbf{x}_k$, which preserves convexity. Since the sum of a strongly convex and a convex function is strongly convex, the cost function $J_k(\mathbf{x}_k)$ defined in (19) is strongly convex. Thus, a unique minimizer exists for the optimization problem in (18). ∎

Having established the well-posedness of the estimator in Lemma 1, we now derive an approximation of the posterior covariance $\mathbf{P}_{k|k}$ based on a second-order expansion of the robust cost function, as formalized in Proposition 2.

Proposition 2. *Suppose that the robust cost function $J_k(\mathbf{x}_k)$, defined in* (19)*, is twice continuously differentiable and strictly convex, and that it admits a unique minimizer $\hat{\mathbf{x}}_{k|k}$. Then, under posterior concentration and smoothness conditions, the posterior covariance can be approximated as*

$$\mathbf{P}_{k|k} \approx \left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1},$$

where $\nabla^2 J_k(\hat{\mathbf{x}}_{k|k}) = \left[\mathbf{P}_{k|k-1}^{-1} + \mathbf{H}_k^\top \mathbf{R}_k^{-\frac{1}{2}} \mathbf{W}_k \mathbf{R}_k^{-\frac{1}{2}} \mathbf{H}_k\right]^{-1}$. *This approximation is justified by the Laplace method, which provides a second-order expansion of the log-posterior around the MAP estimate [39].*

Algorithm 1 Robust Local Filter at node i using the Log-Cosh Loss

1: Initialization: $\mathbf{P}_{i,0|-1} = \mathbf{P}_{i,0}$, $\hat{\mathbf{x}}_{i,0|-1} = \mathbf{x}_{i,0}$

2: for each time step k **do**

a) **Prediction:**

$$\begin{aligned} \hat{\mathbf{x}}_{i,k|k-1} &\leftarrow \mathbf{A}_{k-1}\hat{\mathbf{x}}_{i,k-1|k-1} \\ \mathbf{P}_{i,k|k-1} &\leftarrow \mathbf{A}_{k-1}\mathbf{P}_{i,k-1|k-1}\mathbf{A}_{k-1}^\top + \mathbf{Q}_{k-1} \end{aligned}$$

b) **Robust Measurement Update:**

$$\begin{aligned} \tilde{\mathbf{r}}_{i,k}(\mathbf{x}_{i,k}) &= \mathbf{R}_{i,k}^{-\frac{1}{2}}(\mathbf{z}_{i,k} - \mathbf{H}_{i,k}\mathbf{x}_{i,k}) \\ J_{i,k}(\mathbf{x}_{i,k}) &= \frac{1}{2}(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1})^\top \mathbf{P}_{i,k|k-1}^{-1}(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}) \\ &\quad + \frac{1}{\alpha^2}\sum_{l=1}^{m_i} \log\cosh\left(\alpha\, \tilde{r}_{i,k,l}(\mathbf{x}_{i,k})\right) \\ \hat{\mathbf{x}}_{i,k|k} &\leftarrow \arg\min_{\mathbf{x}_{i,k}} J_{i,k}(\mathbf{x}_{i,k}) \\ \mathbf{W}_{i,k}(\hat{\mathbf{x}}_{i,k|k}) &= \mathrm{diag}\left(\mathrm{sech}^2\left(\alpha\, \tilde{\mathbf{r}}_{i,k}(\hat{\mathbf{x}}_{i,k|k})\right)\right) \\ \nabla^2 J_{i,k}(\hat{\mathbf{x}}_{i,k|k}) &= \mathbf{P}_{i,k|k-1}^{-1} + \mathbf{H}_{i,k}^\top \mathbf{R}_{i,k}^{-\frac{1}{2}} \mathbf{W}_{i,k}(\hat{\mathbf{x}}_{i,k|k}) \mathbf{R}_{i,k}^{-\frac{1}{2}} \mathbf{H}_{i,k} \\ \mathbf{P}_{i,k|k} &\leftarrow \left[\nabla^2 J_{i,k}(\hat{\mathbf{x}}_{i,k|k})\right]^{-1} \end{aligned}$$

3: end for

Proof: This approximation follows from the Laplace method, which approximates the generalized posterior (12) by a Gaussian centered at the MAP estimate. Under the conditions stated in the proposition, the log-posterior is well-approximated by a second-order Taylor expansion. This yields a Gaussian distribution with mean $\hat{\mathbf{x}}_{k|k}$ and covariance $\left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1}$. The expressions for the gradient and Hessian, along with the full justification, appear in Appendix A. ∎

By replacing the standard closed-form Kalman update in (7) with the solution of the convex optimization problem defined in (18), where the cost function (19) incorporates a robust log-cosh loss on the measurement residuals, the proposed filter improves robustness to outliers while preserving the recursive estimation structure. Approximating the generalized posterior by a Gaussian centered at the MAP estimate, with covariance $\left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1}$ as justified in Proposition 2, enables consistent propagation of uncertainty over time.

This local robust filtering mechanism serves as the core component of the distributed estimation algorithm, in which each node independently performs a robust update before engaging in inter-node information fusion. The complete local filtering process at node i is summarized in Algorithm 1, and forms the local step in the distributed estimation scheme presented in the next section.

IV. Fully Distributed Consensus Optimization via EXTRA

In this section, we address the robust estimation problem in a fully distributed network, where each sensor node communicates only with its immediate neighbors to cooperatively

minimize the global objective. At each time step, every node performs the following three operations:

1) **Prediction:** Each node i independently computes its local predicted mean $\hat{\mathbf{x}}_{i,k|k-1}$ and covariance $\mathbf{P}_{i,k|k-1}$, which together form the prior, using the dynamic model in (1) and its previous state estimate.
2) **Distributed Consensus Optimization:** Using its local prior and measurement $\mathbf{z}_{i,k}$, each node participates in a consensus optimization to collaboratively obtain a common global state estimate $\mathbf{x}_k^\star$.
3) **Posterior Covariance Update:** Upon convergence, each node contributes its local Hessian to compute the common posterior covariance $\mathbf{P}_{k|k}$ as the inverse of the global Hessian.

Following a similar procedure discussed in Sec. III-C, to obtain the posterior state estimate in a centralized network with multiple sensors, we construct the corresponding optimization problem by substituting the loss function $J_k(\mathbf{x}_k)$ in (18) with $J_k^c(\mathbf{x}_k)$, defined as follows:

$$\begin{aligned} J_k^c(\mathbf{x}_k) &= \frac{1}{2}\left(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}\right)^\top \mathbf{P}_{k|k-1}^{-1}\left(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}\right) \\ &\quad + \frac{1}{\alpha^2}\sum_{i=1}^{N}\sum_{l=1}^{m_i}\log\cosh\left(\alpha\,\tilde{r}_{i,k,l}(\mathbf{x}_k)\right) \qquad (20) \\ &= \sum_{i=1}^{N} J_{i,k}(\mathbf{x}_k), \qquad (21) \end{aligned}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the global state variable to be estimated, and $\hat{\mathbf{x}}_{k|k-1} \in \mathbb{R}^n$, $\mathbf{P}_{k|k-1} \in \mathbb{R}^{n\times n}$ denote the centralized prior mean and covariance, respectively as defined in Section III-A. The term $J_{i,k}(\mathbf{x}_k)$ represents the local loss function at node i, and is given by

$$\begin{aligned} J_{i,k}(\mathbf{x}_k) &= \frac{1}{2N}\left(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1}\right)^\top \mathbf{P}_{i,k|k-1}^{-1}\left(\mathbf{x}_k - \hat{\mathbf{x}}_{i,k|k-1}\right) \\ &\quad + \frac{1}{\alpha^2}\sum_{l=1}^{m_i}\log\cosh\left(\alpha\,\tilde{r}_{i,k,l}(\mathbf{x}_k)\right), \qquad (22) \end{aligned}$$

where $\hat{\mathbf{x}}_{i,k|k-1}$ is the local predicted mean, $\mathbf{P}_{i,k|k-1}$ is the local predicted prior covariance, and $\tilde{r}_{i,k,l}(\mathbf{x}_k)$ is the l-th component of the local whitened residual vector. This expression assumes that all nodes share a consistent prior mean and covariance, i.e., $\hat{\mathbf{x}}_{i,k|k-1} = \hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}_{i,k|k-1} = \mathbf{P}_{k|k-1}$ for all i. Note that the quadratic term in (22) is scaled by $1/N$ to ensure that, when summed across all nodes, the aggregate prior contribution matches that of the centralized loss in (20). This construction yields a distributed objective in which the centralized prior is correctly incorporated exactly once, while each node contributes its individual measurement likelihood term. Thus, the optimization problem for the posterior state estimate can be equivalently expressed as

$$\min_{\mathbf{x}_k\in\mathbb{R}^n} \quad \sum_{i=1}^{N} J_{i,k}(\mathbf{x}_k), \qquad (23)$$

To enable a fully distributed solution, each node maintains a local copy $\mathbf{x}_{i,k} \in \mathbb{R}^n$ of the state and enforces consensus with its neighbors using equality constraints. These constraints require the local estimates of any two connected nodes to be identical. For a connected network, this ensures that all local estimates across the entire network are driven to a single, common value. The resulting optimization problem is formulated as:

$$\min_{\{\mathbf{x}_{i,k}\}} \sum_{i=1}^{N} J_{i,k}(\mathbf{x}_{i,k}) \quad \text{s.t. } \mathbf{x}_{i,k} = \mathbf{x}_{j,k}, \quad \forall\,(i,j) \in \mathcal{E}. \qquad (24)$$

The optimization problem in (24) recasts the global estimation task as a consensus-constrained optimization, which can be solved using distributed iterative algorithms. Among such methods, we adopt the EXTRA algorithm proposed in [40]. It provides a globally optimal solution using a consensus-corrected gradient method. One of the key advantages of EXTRA is that it guarantees convergence to the centralized robust MAP solution using only local gradient evaluations and neighbor communications. Moreover, unlike the consensus-based Alternating Direction Method of Multipliers (ADMM) [41], which requires each node to solve local optimization subproblems and exchange both primal and dual variables at every iteration, EXTRA eliminates the need for inner-loop optimization and dual variable management. This leads to significantly lower per-iteration computational overhead, while still ensuring convergence. These advantages make it particularly well-suited for large-scale sensor networks in which centralized coordination is infeasible.

In the EXTRA algorithm, two weight matrices are constructed: $W \in \mathbb{R}^{N\times N}$, which is doubly stochastic and compatible with the communication graph, and $\tilde{W} \in \mathbb{R}^{N\times N}$, also symmetric, satisfying the condition

$$\tilde{W} = \frac{1}{2}(\mathbf{I} + W). \qquad (25)$$

The design of the above weight matrices can eliminate the steady-state bias between consensus dynamics and local descent, thereby guaranteeing exact convergence under a fixed step size even when local objectives differ. Next, we present how these two matrices are used in the consensus algorithm to coordinate local and neighborhood updates.

At each consensus iteration ℓ, node i performs the following update:

$$\begin{aligned} \mathbf{x}_{i,k}^{(\ell+1)} &= \mathbf{x}_{i,k}^{(\ell)} + \sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_{j,k}^{(\ell)} - \sum_{j\in\mathcal{N}_i} \tilde{w}_{ij}\mathbf{x}_{j,k}^{(\ell-1)} \\ &\quad - \epsilon\left[\nabla J_{i,k}(\mathbf{x}_{i,k}^{(\ell)}) - \nabla J_{i,k}(\mathbf{x}_{i,k}^{(\ell-1)})\right], \qquad (26) \end{aligned}$$

where w_{ij} and $\tilde{w}_{ij}$ denote the entries of W and $\tilde{W}$, respectively, and $\epsilon > 0$ is a fixed step size. The iteration is initialized by $\mathbf{x}_{i,k}^{(1)} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_{j,k}^{(0)} - \epsilon\nabla J_{i,k}(\mathbf{x}_{i,k}^{(0)})$. Based on (22), the gradient of the local loss is given by

$$\begin{aligned} \nabla J_{i,k}(\mathbf{x}_{i,k}) &= \frac{1}{N}\mathbf{P}_{i,k|k-1}^{-1}(\mathbf{x}_{i,k} - \hat{\mathbf{x}}_{i,k|k-1}) \\ &\quad - \frac{1}{\alpha}\mathbf{H}_{i,k}^\top \mathbf{R}_{i,k}^{-\frac{1}{2}}\boldsymbol{\psi}_{i,k}(\mathbf{x}_{i,k}), \qquad (27) \end{aligned}$$

where $\boldsymbol{\psi}_{i,k}(\mathbf{x}_{i,k}) \triangleq \tanh\left(\alpha\,\tilde{\mathbf{r}}_{i,k}(\mathbf{x}_{i,k})\right)$. The use of the log-cosh function in the local loss (22) ensures that the gradient remains smooth and globally Lipschitz continuous, which guarantees stability of the updates.

Under standard assumptions, including convexity and Lipschitz continuity of $J_{i,k}$, symmetry and connectivity of the graph $\mathcal{G}$, proper initialization, and a sufficiently small step size ϵ (see [40] for the appropriate choice of ϵ), the sequence $\{\mathbf{x}_{i,k}^{(\ell)}\}$ converges to the unique minimizer $\mathbf{x}_k^\star$ of the global objective (24). If the global objective is strongly convex, as ensured by the quadratic prior term in (22), then the convergence rate of EXTRA is linear [40]. Moreover, convergence to a common value across all nodes is guaranteed, i.e.,

$$\lim_{\ell\to\infty} \mathbf{x}_{i,k}^{(\ell)} = \mathbf{x}_k^\star, \quad \forall i \in \mathcal{V}. \tag{28}$$

Since the consensus algorithm we utilized ensures that all nodes share the same estimate $\hat{\mathbf{x}}_{i,k|k} = \mathbf{x}_k^\star$ after convergence, each node adopts it as its local posterior mean. The corresponding posterior covariance is then approximated using the result of Proposition 2, which relates the posterior covariance to the inverse Hessian of the cost evaluated at the MAP estimate. Since the total cost is the sum of local costs as defined in (23), the covariance is given by

$$\mathbf{P}_{k|k} \approx \left[\nabla^2 J_k(\mathbf{x}_k^\star)\right]^{-1} = \left[\sum_{i=1}^{N} \nabla^2 J_{i,k}(\mathbf{x}_k^\star)\right]^{-1}. \tag{29}$$

To compute this sum in a fully distributed manner, each node i initializes $\mathbf{S}_{i,k}^{(0)} = \nabla^2 J_{i,k}(\mathbf{x}_k^\star)$ and iteratively updates it using the weighted averaging rule $\mathbf{S}_{i,k}^{(t+1)} = \sum_{j\in\mathcal{N}_i} a_{ij}\mathbf{S}_{j,k}^{(t)}$, where the weights $a_{ij} \geq 0$ satisfy $\sum_{j\in\mathcal{N}_i} a_{ij} = 1$ and $a_{ij} = 0$ if $j \notin \mathcal{N}_i$. If the communication graph $\mathcal{G}$ is connected and the weight matrix $A = [a_{ij}]$ is symmetric and doubly stochastic, then $\mathbf{S}_{i,k}^{(t)}$ converges to the global average:

$$\lim_{t\to\infty} \mathbf{S}_{i,k}^{(t)} = \frac{1}{N}\sum_{j=1}^{N} \nabla^2 J_{j,k}(\mathbf{x}_k^\star), \tag{30}$$

and since the average consensus protocol [42] returns the mean, each node must multiply the result by N to recover the total sum.[1]

V. Stability

The robust estimator described in Section III attenuates the influence of outliers by down-weighting measurements with large residuals. In this section, we assess the stability properties of the proposed robust estimators by adopting a conservative analytical model, in which residuals exceeding a predefined threshold are treated as uninformative. This abstraction enables the robust filter to be conservatively modeled as a KF with intermittent observations, for which sufficient conditions for mean-square stability can be established according to [45].

The objective is to establish that the expected estimation error covariance remains uniformly bounded over time. Let $\mathbf{P}_{k|k}$ denote the posterior covariance at time k, which approximates the second-order moment of the estimation error,

$$\mathbb{E}\left[(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^\top\right] \approx \mathbf{P}_{k|k}. \tag{31}$$

[1]If the total number of nodes N is not known a priori, it can be estimated in a distributed manner using counting algorithms [43], [44].

We seek to ensure that $\sup_{k\geq 0} \mathbb{E}[\mathbf{P}_{k|k}] < \infty$ under appropriate design of the tuning parameter α. This aligns with the stability condition established in [45, Th. 2].

A. Connection to Intermittent Observations

The robust update derived in Section III-C modifies the standard Kalman filter by introducing a data-dependent weighting matrix in the measurement update. Specifically, as established in Proposition 2, the posterior covariance is approximated as

$$\mathbf{P}_{k|k} \approx \left[\mathbf{P}_{k|k-1}^{-1} + \mathbf{H}_k^\top \mathbf{R}_k^{-\frac{1}{2}} \mathbf{W}_k \mathbf{R}_k^{-\frac{1}{2}} \mathbf{H}_k\right]^{-1}, \tag{32}$$

where $\mathbf{W}_k = \operatorname{diag}\left(\operatorname{sech}^2(\alpha\,\tilde{\mathbf{r}}_k)\right) \in \mathbb{R}^{m\times m}$ is a diagonal matrix that modulates the influence of measurement based on the whitened residuals $\tilde{\mathbf{r}}_k(\mathbf{x}_k)$ (see (14)). When all residual components are small, $\mathbf{W}_k \approx \mathbf{I}$, and the update recovers the structure of the standard KF. For large residuals, the entries in $\mathbf{W}_k$ approach zero, thereby suppressing the influence of the corresponding measurement channels.

This data-dependent weighting introduces a natural thresholding behavior: measurements with small residuals are fully trusted ($\mathbf{W}_k \approx \mathbf{I}$), while those with large residuals are increasingly suppressed. To enable tractable analysis, we conservatively approximate this behavior by introducing a binary observation model: each measurement is either fully incorporated or completely discarded based on a weight threshold $\theta \geq 0.95$, such that only highly reliable residuals are treated as informative. This simplification facilitates stability analysis by aligning the robust filter with the intermittent observation model studied in [45].

Let

$$w_k = \min_{l=1,\ldots,m} [\mathbf{W}_k]_{ll} = \min_{l=1,\ldots,m} \operatorname{sech}^2(\alpha\,\tilde{r}_{k,l}), \tag{33}$$

where $w_k \in [0,1]$ denotes the minimum weight across all measurement components and defines the Bernoulli indicator

$$\gamma_k = \begin{cases} 1, & \text{if } w_k \geq \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{34}$$

which determines whether the measurement at time k is incorporated. The robust filter is then conservatively modeled as a KF with intermittent observations, and the corresponding update step is

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \gamma_k \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}), \tag{35}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \gamma_k \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k|k-1}, \tag{36}$$

where $\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^\top\left(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1}$ is the standard Kalman gain. We denote the resulting posterior covariance under this intermittent update as $\mathbf{P}_{k|k}^{(\text{int})}$, and define the corresponding precision matrix as $\mathbf{\Omega}_{k|k}^{(\text{int})} \triangleq \left(\mathbf{P}_{k|k}^{(\text{int})}\right)^{-1}$. The probability of receiving an informative measurement is thus given by

$$\lambda = \mathbb{P}(\gamma_k = 1) = \mathbb{P}\left(\min_l \operatorname{sech}^2(\alpha\,\tilde{r}_{k,l}) \geq \theta\right), \tag{37}$$

which we refer to as the *inlier probability*. This probability governs the stability of the filter under the intermittent model.

This binary abstraction is conservative, as it neglects the partial information encoded in the nonzero weights of $\mathbf{W}_k$. We now show that the robust filter's posterior precision dominates that of the intermittent filter, ensuring that its uncertainty bound is no worse. This result is formalized in Lemma 2. From (32), the posterior precision of the robust filter satisfies

$$\mathbf{\Omega}_{k|k} = \mathbf{\Omega}_{k|k-1} + \mathbf{H}_k^\top \mathbf{R}_k^{-\frac{1}{2}} \mathbf{W}_k \mathbf{R}_k^{-\frac{1}{2}} \mathbf{H}_k, \tag{38}$$

where $\mathbf{\Omega}_{k|k} \triangleq \mathbf{P}_{k|k}^{-1}$, $\mathbf{\Omega}_{k|k-1} \triangleq \mathbf{P}_{k|k-1}^{-1}$, and $\mathbf{W}_k \succcurlyeq \mathbf{0}$.

Lemma 2 (Covariance Upper Bound via Intermittent KF)**.** *Let $\mathbf{\Omega}_{k|k}$ denote the posterior precision of the robust filter given in* (38)*, and let $\mathbf{\Omega}_{k|k}^{(\mathrm{int})}$ denote the posterior precision of a Kalman filter with intermittent observations updated via* (36)*. Then, under the binary observation model* (34)*, it holds for all k that*

$$\mathbf{\Omega}_{k|k} \succcurlyeq \mathbf{\Omega}_{k|k}^{(\mathrm{int})} \quad \Longrightarrow \quad \mathbf{P}_{k|k} \preccurlyeq \mathbf{P}_{k|k}^{(\mathrm{int})}.$$

Proof: The comparison considers the information contributions of the robust and intermittent filters under the binary observation model (34). When $\gamma_k = 0$, the intermittent model discards the measurement entirely, leading to no update in the posterior precision. In contrast, the robust filter still retains partial information via the nonzero diagonal weights in $\mathbf{W}_k \succcurlyeq \mathbf{0}$, resulting in a strictly greater information gain. When $\gamma_k = 1$, the intermittent model assumes a full measurement update, corresponding to $\mathbf{W}_k = \mathbf{I}$. The robust filter in this case applies a data-dependent weight $\mathbf{W}_k \preccurlyeq \mathbf{I}$, but the weights are guaranteed to exceed a threshold $\theta \geq 0.95$ by construction of the indicator function. Hence, the robust filter still performs an update that is only slightly weaker than the standard Kalman filter. While the robust filter does not dominate in all cases, the inequality $\mathbf{\Omega}_{k|k} \succcurlyeq \mathbf{\Omega}_{k|k}^{(\mathrm{int})}$ holds in the critical case when $\gamma_k = 0$, which governs the worst-case behavior relevant for stability. Since matrix inversion reverses the ordering for positive definite matrices, this implies $\mathbf{P}_{k|k} \preccurlyeq \mathbf{P}_{k|k}^{(\mathrm{int})}$, thereby concluding the proof. ∎

Lemma 2 shows that if the filter remains stable under the conservative assumption that measurements with partial information are fully discarded, the actual robust filter, which incorporates such information, is guaranteed to exhibit equal or superior stability properties.

B. Stability Condition

Under the conservative model introduced in Section V-A, the proposed robust filter can be modeled as a KF with intermittent observations, whose stability is governed by the stochastic Riccati recursion [45, Eq. (15)]. As established in [45, Th. 2], there exists a critical threshold $\lambda_c \in (0, 1)$ for the inlier probability λ such that

$$\lambda > \lambda_c \quad \Longrightarrow \quad \limsup_{k\to\infty} \mathbb{E}[\mathbf{P}_{k|k}] < \infty, \tag{39}$$

whereas for $\lambda < \lambda_c$, the expected error covariance diverges for at least one admissible initial condition. The critical threshold λ_c is bounded within the interval $\underline{\lambda}_c \leq \lambda_c \leq \overline{\lambda}_c$, as established in [45]. The lower bound $\underline{\lambda}_c$ is given by

$$\underline{\lambda}_c = 1 - \frac{1}{\rho(\mathbf{A})^2}, \tag{40}$$

where $\rho(\mathbf{A})$ denotes the spectral radius of the system matrix $\mathbf{A}$. The upper bound $\overline{\lambda}_c$ is obtained by solving a quasi-convex optimization problem involving a linear matrix inequality (LMI), as formalized in [45, Cor. 1]. In the next subsection, we derive a bound on the robustness parameter α that ensures the stability of the proposed robust filter, based on the bounds of the critical threshold λ_c established above.

C. Design of the Robustness Parameter

The robustness parameter α must be chosen to ensure that the inlier probability λ defined in (37) exceeds the critical threshold bounded above by $\overline{\lambda}_c$. Recall that, under the binary observation abstraction introduced in Section V-A, a measurement is considered informative only if all residual components exert sufficient influence on the update. This condition is satisfied when

$$\min_{l=1,\ldots,m} \operatorname{sech}^2(\alpha\, \tilde{r}_{k,l}) \geq \theta. \tag{41}$$

This inequality is equivalent to requiring that the whitened residual vector lies within the symmetric interval:

$$\|\tilde{\mathbf{r}}_k\|_\infty < \delta, \quad \delta = \frac{1}{\alpha} \operatorname{arccosh}\left(\frac{1}{\sqrt{\theta}}\right). \tag{42}$$

Here, δ defines the admissible range of whitened residuals for which measurements are sufficiently reliable to be fully incorporated into the update. Evaluating the probability that condition (42) holds requires knowledge of the residual distribution, which is unknown and assumed to be non-Gaussian. We focus on heavy-tailed distributions that are unimodal and symmetric, such as the Student's t-distribution, ensuring that the inlier region is centered and the single-threshold condition (42) remains meaningful. This makes a closed-form solution for α under the true noise model intractable. To address this, we adopt a two-step robust design strategy. First, for analytical tractability, we approximate the whitened residual as $\tilde{\mathbf{r}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_r)$, yielding a nominal inlier probability λ_G. This Gaussian surrogate is not intended to model the true distribution but to provide a tractable baseline for bounding performance. Second, to account for the mismatch introduced by heavy-tailed behavior, we introduce a safety margin $\Delta > 0$. Since large residuals occur more frequently under such distributions, the true inlier probability may fall below λ_G. Enforcing the stricter condition $\lambda_G \geq \overline{\lambda}_c + \Delta$ ensures stability even when the residual distribution deviates significantly from the Gaussian assumption.

The value of Δ can be determined from empirical data. One approach is non-parametric: estimate the actual inlier probability λ_{HT} directly from a representative dataset of observed residuals, and set $\Delta = \lambda_G - \lambda_{\mathrm{HT}}$. Alternatively, a parametric model, such as a Student's t-distribution, can be fitted to the residuals to quantify the deviation from the nominal Gaussian approximation. Both strategies provide data-driven estimates

of the tail behavior without altering the analytical derivation of the stability condition. To evaluate whether the Gaussian surrogate inlier probability λ_G satisfies the stability condition $\lambda_G \geq \overline{\lambda}_c + \Delta$, we derive a conservative lower bound on λ_G, as stated in Proposition 3.

Proposition 3 (Conservative Bound on Inlier Probability)**.** *Let the whitened residual vector be approximated as* $\tilde{\mathbf{r}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_r)$*, where* $\mathbf{\Sigma}_r = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{H}_k\overline{\mathbf{P}}\mathbf{H}_k^\top + \mathbf{R}_k)\mathbf{R}_k^{-\frac{1}{2}}$*. Then, for any threshold* $\delta > 0$*, the inlier probability satisfies*

$$\lambda_G \equiv \mathbb{P}\left(\|\tilde{\mathbf{r}}_k\|_\infty < \delta\right) \geq \left[2\Phi\left(\delta/\sigma_\rho\right) - 1\right]^m, \tag{43}$$

where m *is the measurement dimension,* $\Phi(\cdot)$ *is the standard normal cumulative distribution function, and* $\sigma_\rho = \sqrt{\rho(\mathbf{\Sigma}_r)}$ *denotes the upper bound on the marginal standard deviations, derived from the spectral radius of* $\mathbf{\Sigma}_r$*.*

Proof: Under the Gaussian surrogate model, the whitened residual $\tilde{\mathbf{r}}_k$ is derived as follows. Based on the measurement model in (2), and assuming a linear-Gaussian approximation for the prediction error $\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_{k|k-1})$ and measurement noise $\boldsymbol{\nu}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$, the residual is

$$\mathbf{r}_k = \mathbf{H}_k(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) + \boldsymbol{\nu}_k,$$

which follows a Gaussian distribution:

$$\mathbf{r}_k \sim \mathcal{N}(\mathbf{0},\, \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k).$$

The whitened residual is $\tilde{\mathbf{r}}_k = \mathbf{R}_k^{-\frac{1}{2}}\mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_r)$, where

$$\mathbf{\Sigma}_r = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k)\mathbf{R}_k^{-\frac{1}{2}}.$$

For analyzing asymptotic stability, we replace $\mathbf{P}_{k|k-1}$ with the steady-state solution $\overline{\mathbf{P}}$ of the discrete-time algebraic Riccati equation (DARE), as is standard in Kalman filter stability studies (e.g., [45]) to capture long-term boundedness; this yields

$$\mathbf{\Sigma}_r = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{H}_k\overline{\mathbf{P}}\mathbf{H}_k^\top + \mathbf{R}_k)\mathbf{R}_k^{-\frac{1}{2}}. \tag{44}$$

For each component $\tilde{r}_{k,l} \sim \mathcal{N}(0, \sigma_l^2)$, where σ_l^2 is the l-th diagonal element of $\mathbf{\Sigma}_r$, the marginal inlier probability is given by

$$\mathbb{P}(|\tilde{r}_{k,l}| < \delta) = 2\Phi(\delta/\sigma_l) - 1.$$

Since the spectral radius $\rho(\mathbf{\Sigma}_r)$ (largest eigenvalue) upper-bounds the diagonal elements of $\mathbf{\Sigma}_r$, it follows that $\sigma_l \leq \sigma_\rho$ for all l, and thus

$$\mathbb{P}(|\tilde{r}_{k,l}| < \delta) \geq 2\Phi(\delta/\sigma_\rho) - 1 \triangleq p_\rho.$$

Since $\|\tilde{\mathbf{r}}_k\|_\infty < \delta$ is equivalent to $|\tilde{r}_{k,l}| < \delta$ for all $l = 1, \ldots, m$, Šidák's inequality [46] implies

$$\mathbb{P}\left(\|\tilde{\mathbf{r}}_k\|_\infty < \delta\right) \geq \prod_{l=1}^{m} \mathbb{P}(|\tilde{r}_{k,l}| < \delta) \geq p_\rho^m.$$

Substituting back yields the conservative bound in (43). ■

The use of a common upper bound σ_ρ and the application of Šidák's inequality ensure that the result remains valid even when the components of $\tilde{\mathbf{r}}_k$ are correlated. This construction is conservative, as correlation between residual components can only increase the joint inlier probability relative to the product of the marginals [46]. Combining the robust stability condition $\lambda_G \geq \overline{\lambda}_c + \Delta$ with the conservative bound from Proposition 3 yields a tractable condition on α that ensures filter stability. This result is formalized in Theorem 1.

Theorem 1 (Sufficient Condition for Stability)**.** *Let the system satisfy the conditions of [45, Th. 2]: 1)* $(\mathbf{A}, \mathbf{Q}^{1/2})$ *is controllable, 2)* $(\mathbf{A}, \mathbf{H})$ *is detectable, and 3)* $\mathbf{A}$ *is unstable. Let* $\Delta > 0$ *denote a user-defined safety margin that accounts for deviations of the true residual distribution from the Gaussian surrogate. If the robustness parameter* α *satisfies*

$$0 < \alpha \leq \frac{\operatorname{arccosh}\left(\theta^{-1/2}\right)}{\sigma_\rho\, \Phi^{-1}\left(\frac{(\overline{\lambda}_c+\Delta)^{1/m}+1}{2}\right)}, \tag{45}$$

where θ*,* σ_ρ*, and* $\overline{\lambda}_c$ *are defined in* (41)*,* (43)*, and Section V-B, respectively, then the expected estimation error covariance of the robust filter remains uniformly bounded, i.e.,* $\sup_{k\geq 0} \mathbb{E}\left[\mathbf{P}_{k|k}\right] < \infty$*.*

Proof: The result follows by applying the stability condition $\lambda > \overline{\lambda}_c$ from [45, Th. 2], using the conservative bound on λ_G from Proposition 3. A detailed derivation is provided in Appendix B. ■

D. Extension to the Distributed Setting

The stability result for the centralized robust filter carries over to the fully-distributed implementation because of the convergence properties of the consensus algorithm. In the ideal case where the consensus phase is run to completion, all local estimates converge to the exact global minimizer of the aggregate cost function. At this point, every node holds the same robust MAP estimate and posterior covariance, making the ensuing prediction stage algebraically identical to that of the centralized filter whose stability is already established.

More practically, when only a finite number of consensus iterations are performed, a small disagreement error between nodes persists. This residual error can be modeled as a bounded, zero-mean perturbation that enters the prediction step as a small, additive process noise term, resulting in a bounded inflation of the error covariance. Since the centralized filter is proven to be mean-square stable, the introduction of this controllable perturbation does not alter this fundamental property.

Hence, as long as the communication graph is connected and each sampling window allows for a sufficient number of consensus iterations, the distributed estimator inherits the mean-square stability guaranteed by the centralized analysis.

E. Adaptive Robustness Parameter

In the previous subsection, we derived the range of the robustness parameter α that satisfies a conservative stability requirement. Specifically, an upper bound on α was obtained to ensure that the inlier probability λ_G exceeds the critical threshold $\overline{\lambda}_c$, thereby guaranteeing bounded estimation error. However, this design yields a static α that is applied uniformly across all time steps and residuals. Although such a choice

guarantees stability, it may not be optimal for performance in the presence of heavy-tailed noise.

According to (42), the inlier threshold $\delta = \operatorname{arccosh}(1/\sqrt{\theta})/\alpha$ increases as α decreases, thereby expanding the region $\|\tilde{\mathbf{r}}_k\|_\infty < \delta$ where residuals are assigned significant weight by the function $\operatorname{sech}^2(\alpha\tilde{r}_k)$. As a result, the filter becomes less selective and admits even moderately large residuals as informative, leading to behavior similar to a standard KF. This makes the filter more vulnerable to degradation in non-Gaussian or outlier-prone settings. Conversely, increasing α narrows the inlier region and strengthens outlier suppression, but reduces the probability that residuals fall within this region, potentially causing λ_G to fall below $\overline{\lambda}_c$, thereby compromising stability.

This trade-off motivates a dynamic strategy in which the robustness parameter α_k is adapted as a function of the residual magnitude at time k. The objective is to maintain a small value of α_k within the inlier region, thereby assigning greater weight to informative measurements while preserving stability, and to progressively increase α_k outside this region to attenuate the influence of large residuals more effectively. Although the conservative analysis presented in Section V-A models residuals exceeding the inlier threshold δ as uninformative and discards them via a binary abstraction, in practice such residuals continue to contribute to the update due to the smooth decay of the weighting function $\operatorname{sech}^2(\alpha\tilde{r}_k)$, particularly when α is small. Hence, adaptively increasing α_k outside the inlier region provides a principled means of capturing partial information from moderate outliers while effectively suppressing the impact of extreme deviations, all without compromising the inlier probability λ_G that ensures filter stability.

To implement this trade-off, we define a smooth, monotonic adaptation of the robustness parameter α_k based on the scaled magnitude of the whitened residual. Let $\alpha_{\min}$ denote the largest value that still satisfies the stability bound (45); it serves as the baseline value in the adaptive schedule to ensure the inlier probability λ_G exceeds the critical threshold $\overline{\lambda}_c$

$$\alpha_k = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \frac{1}{1 + \exp\left[-\beta\left(\frac{\|\tilde{\mathbf{r}}_k\|_2}{\sigma_\rho} - c\right)\right]}, \quad (46)$$

where $\alpha_{\max} > \alpha_{\min}$ is chosen to ensure effective outlier attenuation, $\beta > 0$ controls the steepness of the transition and σ_ρ is defined in (43). The parameter $c > 0$ defines the midpoint of the logistic transition and should satisfy $c > \delta/\sigma_\rho$ to ensure that the inflection point lies outside the inlier region, where δ is defined in (42).

The scaled residual norm $\frac{\|\tilde{\mathbf{r}}_k\|_2}{\sigma_\rho}$, provides a normalized, dimensionless and monotonically increasing measure of the deviation between the measurement and prediction, which adaptively adjusts the influence of each measurement based on its magnitude of deviation. For small residuals within the inlier region, the term $\exp\left[-\beta\left(\frac{\|\tilde{\mathbf{r}}_k\|_2}{\sigma_\rho} - c\right)\right]$ in (46) tends to 0, resulting in $\alpha_k \approx \alpha_{\min}$. This ensures that informative measurements retain high influence while preserving the inlier probability λ_G. As the residual magnitude increases, the term $\exp\left[-\beta\left(\frac{\|\tilde{\mathbf{r}}_k\|_2}{\sigma_\rho} - c\right)\right]$ in (46) decreases, and consequently, α_k increases gradually and approaches $\alpha_{\max}$ in the presence of extreme outliers. This allows stronger rejection of extreme outliers while incorporating information from moderate deviations. The logistic formulation of α_k provides a smooth and differentiable transition well-suited for recursive filtering and iterative optimization. The logistic schedule in (46) therefore produces a measurement-dependent robustness parameter whose value is (i) guaranteed not to violate the stability bound and (ii) progressively larger for increasingly atypical residuals.

To visualise the effect, Fig. 1 compare the weighting behavior of $\operatorname{sech}^2(\alpha_k\tilde{r}_k)$ under both adaptive and fixed values of α. It illustrates the distribution of weights applied to whitened residuals $(\tilde{r}_k)$. The adaptive scheme allocates higher weights within the inlier region, while both strategies assign negligible weights to large residuals, effectively suppressing outliers.

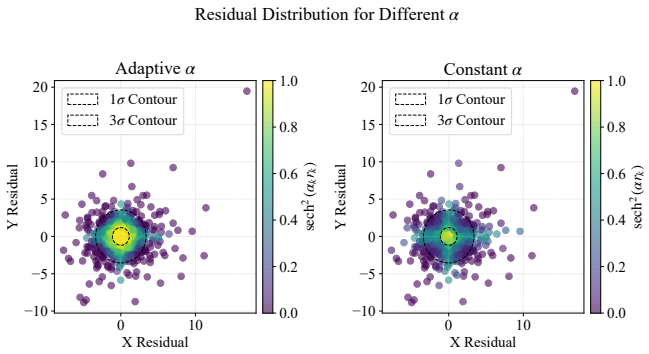

Fig. 1. Whitened residuals weighted by $\operatorname{sech}^2(\alpha_k\tilde{r}_k)$ under adaptive and constant α strategies. Color intensity reflects the weight assigned to each residual, with lower weights indicating down-weighted outliers. Dashed curves denote the 1σ and 3σ standard deviation contours of the residual distribution.

VI. Results

We evaluate the proposed algorithm using the linear dynamical model described in Section II, applied to a two-dimensional target tracking scenario involving a network of 50 sensor nodes, whose communication topology is illustrated in Fig. 2. The global state at time step k is defined as $\mathbf{x}_k = \begin{bmatrix} x_k & y_k & \dot{x}_k & \dot{y}_k \end{bmatrix}^\top$, where (x_k, y_k) and $(\dot{x}_k, \dot{y}_k)$ denote the position and velocity components, respectively. The state evolves according to a time-invariant constant velocity model with unknown accelerations, and the sampling interval is fixed at $\Delta T = 0.2$. The state transition matrix and process noise covariance are instantiated as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{Q} = \sigma_a^2 \begin{bmatrix} \frac{1}{4}\Delta T^4 & 0 & \frac{1}{2}\Delta T^3 & 0 \\ 0 & \frac{1}{4}\Delta T^4 & 0 & \frac{1}{2}\Delta T^3 \\ \frac{1}{2}\Delta T^3 & 0 & \Delta T^2 & 0 \\ 0 & \frac{1}{2}\Delta T^3 & 0 & \Delta T^2 \end{bmatrix},$$

where σ_a^2 denotes the variance of the zero-mean, independent acceleration components $\ddot{x}$ and $\ddot{y}$, and the process noise is

distributed as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$. Each sensor node observes only the position, with an observation matrix

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The measurement noise $\boldsymbol{\nu}_i \in \mathbb{R}^2$ is modeled as a zero-mean multivariate Student-t random vector:

$$\boldsymbol{\nu}_i \sim \mathcal{T}_\eta(\mathbf{0}, \mathbf{R}_i), \tag{47}$$

where $\eta > 2$ denotes the degrees of freedom and $\mathbf{R}_i \in \mathbb{R}^{2\times 2}$ is a positive definite scale matrix. This distribution admits the equivalent Gaussian scale mixture representation:

$$\begin{aligned} \boldsymbol{\nu}_i &= \frac{\boldsymbol{\xi}_i}{\sqrt{w}}, \\ \boldsymbol{\xi}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i), \quad w \sim \text{Gamma}\left(\frac{\eta}{2}, \frac{\eta}{2}\right). \end{aligned} \tag{48}$$

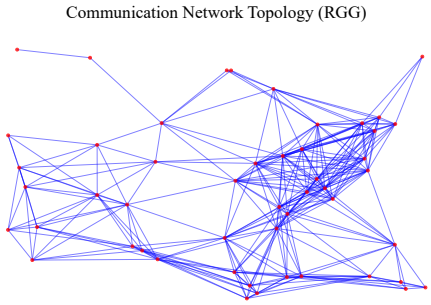

Fig. 2. Network topology with 50 nodes used in the simulation experiments.

To assess robustness of the proposed algorithm under varying sensing conditions, including different percentages of sensor nodes corrupted by heavy-tailed Student's t-distributed noise and the presence of external disturbances[2] in addition to measurement corruption, by analyzing the Root Mean Square Error (RMSE) of position estimates. Fig. 3 shows the RMSE distributions for a scenario where 25% of the sensor nodes are affected by heavy-tailed Student's t-distributed noise (47). We compare the performance of the proposed log-cosh-based filters (both centralized and distributed) with the standard centralized KF and the IMQ filters (both centralized and distributed). The IMQ filters, based on a covariance inflation strategy proposed in [26], are implemented using the Information-weighted Consensus Filter (ICF) framework [47]. The distributed version of the proposed log-cosh filter is implemented using the EXTRA algorithm described in Section IV. All methods are evaluated over 100 Monte Carlo runs, with each RMSE value representing the average estimation error across all nodes for a given run. We can observed from Fig. 3 that the proposed filters (centralized and distributed) outperform both the IMQ filters and the standard KF, with the centralized one achieving the lowest error due to access to complete information. Table I summarizes the average estimation errors under four different outlier contamination levels, demonstrating a consistent 5% to 10% reduction in RMSE compared to the IMQ filter.

[2]External disturbances refer to unmodeled state perturbations that directly affect the target dynamics, such as abrupt changes in velocity or trajectory, and are distinct from measurement noise corruption, which affects the observations made by the sensor nodes.

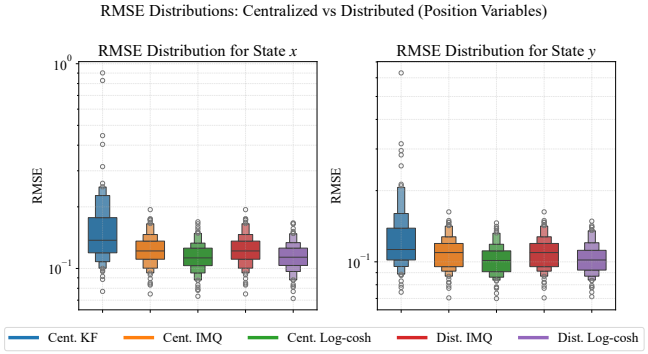

Fig. 3. Log-scale distribution of position RMSE across 100 simulation runs, with 25% of the sensor nodes corrupted by heavy-tailed Student's t-distributed noise.

TABLE I
AVERAGE RMSE IN x AND y POSITION ESTIMATES ACROSS 100 SIMULATION RUNS FOR THE DISTRIBUTED IMQ AND LOG-COSH FILTERS UNDER VARYING PERCENTAGES OF OUTLIERS.

% Outliers	Distributed IMQ		Distributed Log-cosh	
	RMSE-x	RMSE-y	RMSE-x	RMSE-y
25%	0.1249	0.1094	0.1148	0.1026
50%	0.1274	0.1136	0.1211	0.1094
75%	0.1383	0.1189	0.1298	0.1142
100%	0.1406	0.1195	0.1313	0.1145

To assess performance under more challenging conditions, we introduce random external disturbances that perturb the position dynamics in addition to the heavy-tailed Student's t-distributed measurement noise. Specifically, at each time step, with a small fixed probability, a disturbance vector is added to the position component of the state. This vector has a random direction, and its magnitude is sampled uniformly up to a predefined maximum, simulating sudden jumps or maneuvers in the target's motion. The resulting RMSE distributions are shown in Fig. 4. The degradation observed in the IMQ filter under these conditions is substantially mitigated by the proposed log-cosh-based method, which achieves up to 20% lower RMSE. Table II provides the average error metrics under this combined disturbance regime. The performance improvements of the proposed log-cosh-based filter are attributed to the smooth saturating nature of the log-cosh loss, which down-weights large residuals without entirely discarding their contribution, thereby retaining useful information from moderately informative deviations while attenuating extreme outliers. This property enhances the filter's resilience to both observation anomalies and abrupt process noise.

Howover, beyond estimation accuracy, the practical utility of a distributed algorithm hinges on its communication efficiency and the quality of inter-node consensus. Fig. 5 presents the average number of consensus iterations required per simulation

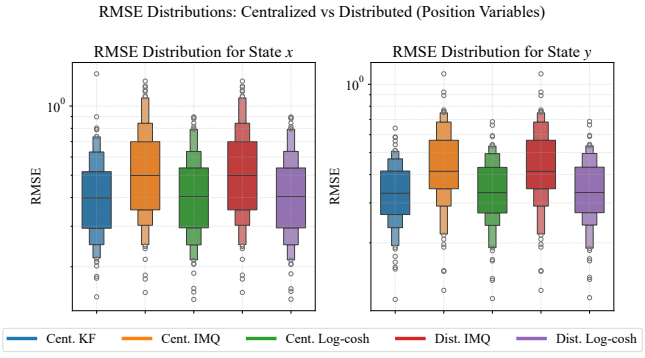

Fig. 4. Log-scale distribution of position RMSE across 100 simulation runs, with 25% of the sensor nodes corrupted by heavy-tailed Student's t-distributed noise and external disturbances.

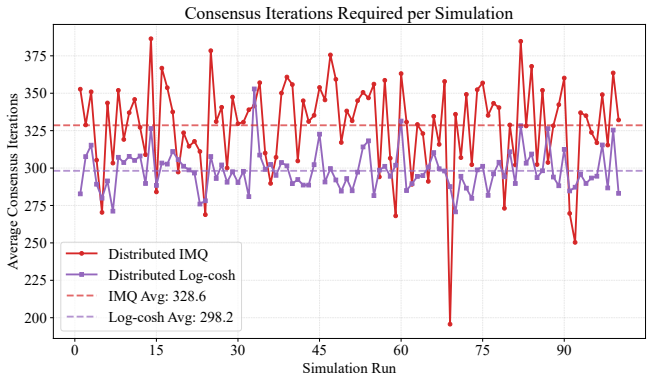

Fig. 5. Average number of consensus iterations per simulation run for the distributed IMQ and log-cosh-based filters under heavy-tailed measurement noise, where 25% of the sensor nodes are corrupted.

TABLE II
AVERAGE RMSE IN x AND y POSITION ESTIMATES ACROSS 100 SIMULATION RUNS FOR THE DISTRIBUTED IMQ AND LOG-COSH FILTERS UNDER VARYING PERCENTAGES OF OUTLIERS WITH PROCESS DISTURBANCES.

% Outliers	Distributed IMQ		Distributed Log-cosh	
	RMSE-x	RMSE-y	RMSE-x	RMSE-y
25%	0.5260	0.4480	0.4319	0.3591
50%	0.4975	0.3865	0.4336	0.3356
75%	0.5119	0.4402	0.4451	0.3778
100%	0.5705	0.4888	0.4884	0.4171

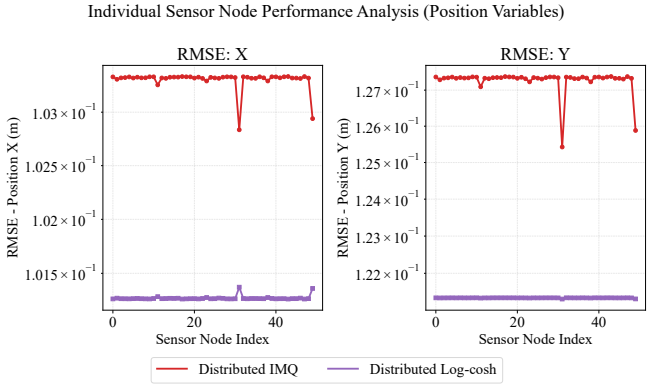

Fig. 6. Node-wise RMSE values for the distributed IMQ and log-cosh-based filters under heavy-tailed measurement noise, with 25% of the sensor nodes corrupted, in a single simulation run.

run for the distributed IMQ and log-cosh-based filters under heavy-tailed Student's t-distributed noise, with a consensus tolerance of 10^{-2} used as the stopping criterion. The proposed distributed log-cosh-based filter, implemented via the EXTRA algorithm, consistently converges with fewer iterations than the IMQ filter implemented within the ICF framework. This performance gain is attributed to the combined strengths of the EXTRA algorithm, which ensures fast convergence in distributed optimization, and the designed log-cosh loss, whose smooth and stable gradients enable reliable update steps. Fig. 6 demonstrates the consistency of estimation performance across individual sensor nodes for a representative simulation with 25% of nodes affected by outliers. The distributed log-cosh-based filter exhibits uniformly low RMSE across all sensor nodes, indicating effective suppression of outlier influence and consistent agreement among local estimates. In contrast, the distributed IMQ filter shows larger variation in node-wise errors, reflecting its sensitivity to measurement corruption and less reliable consensus. These results highlight the robustness of the proposed approach and its ability to maintain accurate and coherent estimates across the network.

Fig. 7 and Fig. 8 compare the estimated trajectories from different algorithms against the true system trajectory under two representative simulation scenarios. In Fig. 7, 25% of the sensor nodes are affected by Student's t-distributed measurement noise (47). Under this setting, the standard centralized KF exhibits noticeable deviations and fails to accurately track the true trajectory, whereas both centralized and distributed versions of the proposed log-cosh-based filter demonstrate significantly improved tracking performance and closely follow the ground truth. Fig. 8 depicts results under the more challenging scenario where random external disturbances are introduced in addition to the 25% outlier-contaminated measurements. In this case, the proposed filter continues to outperform the IMQ filter in both centralized and distributed configurations, maintaining a tighter track of the true trajectory. These results highlight the robustness of the proposed method in the presence of both measurement anomalies and process disturbances.

Fig. 9 compares the RMSE distributions for the adaptive and constant α strategies, revealing improvements of less than 5%, which are not substantial enough to conclusively establish the superiority of the adaptive approach. This limited performance gain may be a consequence of the difficulty in tuning the logistic adjustment parameters in (46), suggesting that the observed results reflect the sensitivity of the method to parameter selection rather than its intrinsic limitations. This highlights that while adaptive strategies are promising, their practical deployment requires careful consideration of hyperparameter tuning, which remains an important direction for future work.

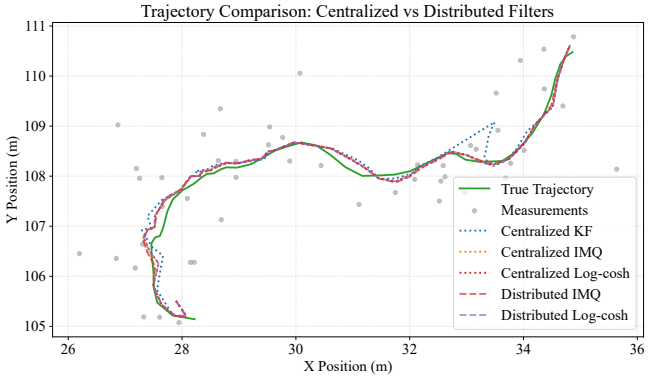

Fig. 7. Estimated trajectories obtained from different algorithms compared against the true trajectory, with 25% of the sensor nodes corrupted by heavy-tailed measurement noise.

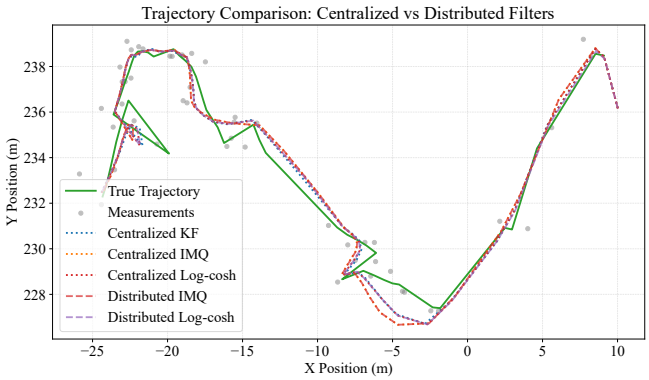

Fig. 8. Estimated trajectories obtained from different algorithms compared against the true trajectory, with 25% of the sensor nodes corrupted by heavy-tailed measurement noise and additional external disturbances affecting the process dynamics.

VII. Conclusion

This paper developed and validated a robust, fully distributed state estimation framework capable of operating effectively under unimodal, symmetric heavy-tailed noise distributions, which model the presence of outliers. By embedding a smooth, convex log-cosh loss into a generalized Bayesian inference formulation, the proposed method effectively limits the influence of large residuals while preserving the differentiability necessary for efficient gradient-based optimization. An analytical stability analysis is presented, along with the derivation of a robustness parameter regime that guarantees stability.

The proposed estimator was implemented in a distributed setting using the EXTRA algorithm, enabling exact convergence through local message exchanges and fixed step size. Numerical evaluations in a target tracking scenario demonstrated consistent performance gains over both the standard Kalman filter and the IMQ-based robust filters. In particular, the proposed method achieved up to 20% lower RMSE under heavy-tailed Student's t-distributed measurement noise and external disturbances, with fewer consensus iterations required for convergence. To improve adaptability, a dynamic adjustment rule was developed for the robustness parameter.

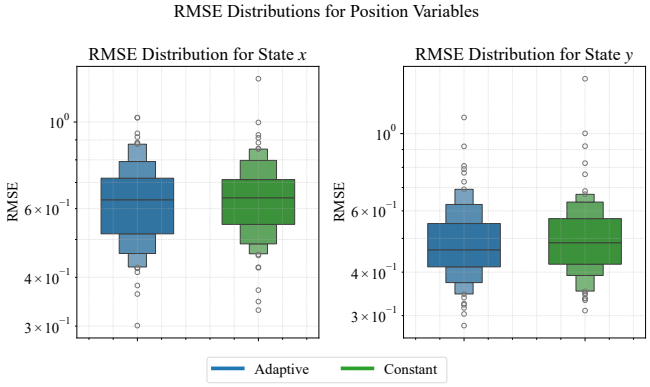

Fig. 9. Log-scale distribution of position RMSE across 100 simulation runs for adaptive and constant α strategies, with 25% of the sensor nodes corrupted by heavy-tailed Student's t-distributed noise.

Overall, this work provides a theoretically grounded, computationally efficient, and empirically validated framework for resilient state estimation, offering a practical solution for modern distributed sensor networks operating in adversarial environments.

Future work could be a further investigation into more sophisticated adaptive mechanisms for robustness parameter tuning, such as those based on online hyperparameter tuning, to fully unlock their potential. We are also planning to extend the current framework to scenarios with non-static network topologies or time-varying noise models.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Commun. Mag.*, vol. 40, pp. 102–114, Aug. 2002.

[2] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy, and M. S. Shehata, "Structural health monitoring using wireless sensor networks: a comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 19, pp. 1403–1423, 2017.

[3] J. Sijs, M. Lazar, P. P. J. v. d. Bosch, and Z. Papp, "An overview of non-centralized Kalman filters," in *Proc. IEEE Int. Conf. Control Appl.* IEEE, Jul. 2008, pp. 738–744.

[4] R. Abdolee and B. Champagne, "Centralized adaptation for parameter estimation over wireless sensor networks," *IEEE Commun. Lett.*, vol. 19, pp. 1624–1627, Jul. 2015.

[5] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*, 20th ed., S. M. Kay, Ed. USA: Prentice-Hall, Inc., 1993, vol. 1.

[6] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Comput. Netw.*, vol. 52, pp. 2292–2330, Aug. 2008.

[7] M. Hassan, G. Salut, M. Singh, and A. Titli, "A decentralized computational algorithm for the global Kalman filter," *IEEE Trans. Autom. Control*, vol. 23, pp. 262–268, Apr. 1978.

[8] J. Speyer, "Computation and transmission requirements for a decentralized linear-quadratic-Gaussian control problem," *IEEE Trans. Autom. Control*, vol. 24, pp. 266–269, Apr. 1979.

[9] L. Shi, K. H. Johansson, and R. M. Murray, "Estimation over wireless sensor networks: tradeoff between communication, computation and estimation qualities," in *IFAC Proc. Vol.*, vol. 41. Elsevier, 2008, pp. 605–611.

[10] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," in *Proc. Ieee*, vol. 95. IEEE, Jan. 2007, pp. 215–233.

[11] R. Olfati-Saber, "Distributed Kalman filtering for sensor networks," in *Proc. IEEE Conf. Decis. Control*. IEEE, 2007, pp. 5492–5498.

[12] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, pp. 2069–2084, Sep. 2010.

[13] A. G. Dimakis, A. D. Sarwate, and M. J. Wainwright, "Geographic gossip: efficient aggregation for sensor networks," in *Proc. Int. Conf. Inf. Process. Sens. Netw. (IPSN)*. New York, NY: ACM, 2006, pp. 69–76.

[14] R. Olfati-Saber, "Distributed Kalman filter with embedded consensus filters," in *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, vol. 2005. [Piscataway, NJ]: IEEE, 2005, pp. 8179–8184.

[15] R. Olfati-Saber and J. S. Shamma, "Consensus filters for sensor networks and distributed sensor fusion," in *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, vol. 2005. [Piscataway, NJ]: IEEE, 2005, pp. 6698–6703.

[16] A. T. Kamal, C. Ding, B. Song, J. A. Farrell, and A. K. Roy-Chowdhury, "A generalized Kalman consensus filter for wide-area video networks," in *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*. IEEE, Dec. 2011, pp. 7863–7869.

[17] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proc. Am. Control Conf.*, vol. 4. IEEE, 1997, pp. 2369–2373.

[18] G. Battistelli, L. Chisci, G. Mugnai, A. Farina, and A. Graziano, "Consensus-based linear and nonlinear filtering," *IEEE Trans. Autom. Control*, vol. 60, pp. 1410–1415, May 2015.

[19] S. P. Talebi, S. Kanna, Y. Xia, and D. P. Mandic, "Cost-effective diffusion Kalman filtering with implicit measurement exchanges," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, Mar. 2017, pp. 4411–4415.

[20] J. Qin, J. Wang, L. Shi, and Y. Kang, "Randomized consensus-based distributed Kalman filtering over wireless sensor networks," *IEEE Trans. Autom. Control*, vol. 66, pp. 3794–3801, Jul. 2021.

[21] V. Barnett and T. Lewis, *Outliers in statistical data*, 3rd ed., ser. Wiley series in probability and mathematical statistics. Chichester [u.a.]: John Wiley & Sons, 1994.

[22] H. Zhu, M. J. V. Amuri, X. Li, and J. Shen, "Mean-shift-based outliers-robust distributed Kalman filter for wireless sensor network systems," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–12, 2025.

[23] M. Roth, T. Ardeshiri, E. Özkan, and F. K. Gustafsson, "Robust Bayesian filtering and smoothing using Student's t distribution," arXiv, 2017.

[24] P. Dong, Z. Jing, H. Leung, K. Shen, and M. Li, "Robust consensus nonlinear information filter for distributed sensor networks with measurement outliers," *IEEE Trans. Cybern.*, vol. 49, pp. 3731–3743, Oct. 2019.

[25] S. Modalavalsa, U. K. Sahoo, A. K. Sahoo, and S. Kumar, "Diffusion minimum generalized rank norm over distributed adaptive networks: formulation and performance analysis," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, pp. 669–683, Dec. 2019.

[26] G. Duran-Martin *et al.*, "Outlier-robust Kalman filtering through generalised Bayes," in *Proc. Int. Conf. Mach. Learn.*, R. Salakhutdinov *et al.*, Eds., vol. 235. PMLR, Jul. 2024, pp. 12 138–12 171.

[27] C. Ding and B. Jiang, "L1-norm error function robustness and outlier regularization," May 2017.

[28] K. Gokcesu and H. Gokcesu, "Generalized huber loss for robust learning and its efficient minimization for a robust statistics," Aug. 2021.

[29] M. O. Sayin, N. D. Vanli, and S. S. Kozat, "A novel family of adaptive filtering algorithms based on the logarithmic cost," *IEEE Trans. Signal Process.*, vol. 62, pp. 4411–4424, Sep. 2014.

[30] R. Izanloo, S. A. Fakoorian, H. S. Yazdi, and D. Simon, "Kalman filtering based on the maximum correntropy criterion in the presence of non-Gaussian noise," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*. IEEE, Mar. 2016, pp. 500–505.

[31] B. Chen, X. Liu, H. Zhao, and J. C. Príncipe, "Maximum correntropy kalman filter," arXiv preprint arXiv:1509.04580, 2015.

[32] B. Chen, L. Xing, B. Xu, H. Zhao, N. Zheng, and J. C. Principe, "Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering," *IEEE Trans. Signal Process.*, vol. 65, pp. 2888–2901, Jul. 2017.

[33] C. Liu and M. Jiang, "Robust adaptive filter with lncosh cost," *Signal Process.*, vol. 168, p. 107348, Mar. 2020.

[34] S. Wang, W. Wang, K. Xiong, H. H. C. Iu, and C. K. Tse, "Logarithmic hyperbolic cosine adaptive filter and its performance analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, pp. 2512–2524, Apr. 2021.

[35] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslov. Math. J.*, vol. 23, pp. 298–305, 1973.

[36] P. G. Bissiri, C. C. Holmes, and S. G. Walker, "A general framework for updating belief distributions," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 78, no. 5, pp. 1103–1130, 02 2016. [Online]. Available: https://doi.org/10.1111/rssb.12158

[37] R. A. Saleh and A. K. M. E. Saleh, "Statistical properties of the log-cosh loss function used in machine learning," Aug. 2022.

[38] A. Jeendgar, T. Devale, S. S. Dhavala, and S. Saha, "Loggene: a smooth alternative to check loss for deep healthcare inference tasks," Jun. 2022.

[39] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Am. Stat. Assoc.*, vol. 81, pp. 82–86, Mar. 1986.

[40] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," arXiv preprint arXiv:1404.6264, 2014.

[41] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers, 2011, vol. 3, no. 1.

[42] K. Avrachenkov, M. E. Chamie, and G. Neglia, "A local average consensus algorithm for wireless sensor networks," in *Proc. Int. Conf. Distrib. Comput. Sens. Syst. Workshops (DCOSS)*. Piscataway, NJ: IEEE, 2011.

[43] R. v. D. Bovenkamp, F. Kuipers, and P. V. Mieghem, "Gossip-based counting in dynamic networks," in *Lect. Notes Comput. Sci.*, R. Bestak, L. Kencl, L. E. Li, J. Widmer, and H. Yin, Eds., vol. LNCS-7290. Springer, Jul. 2012, pp. 404–417.

[44] A. Saha, K. Meel, S. Roy, and A. V. Vullikanti, "Memory and communication efficient algorithm for decentralized counting of nodes in networks," *PLoS ONE*, vol. 16, p. 0259736, Nov. 2021.

[45] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Autom. Control*, vol. 49, pp. 1453–1464, Sep. 2004.

[46] Z. Sidak, "Rectangular confidence regions for the means of multivariate normal distributions," *J. Amer. Statist. Assoc.*, vol. 62, no. 318, p. 626, Jun. 1967.

[47] A. T. Kamal, J. A. Farrell, and A. K. Roy-Chowdhury, "Information weighted consensus," in *Proc. IEEE Conf. Decis. Control (CDC)*. IEEE, Dec. 2012, pp. 2732–2737.

APPENDIX A
PROOF OF PROPOSITION 2

Consider the specific loss function in (19), defined as

$$J_k(\mathbf{x}_k) = \frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^\top \mathbf{P}_{k|k-1}^{-1}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) + \frac{1}{\alpha^2}\sum_{l=1}^{m} \log\cosh\left(\alpha\, \tilde{r}_{k,l}(\mathbf{x}_k)\right), \tag{49}$$

where $\tilde{\mathbf{r}}_k(\mathbf{x}_k) = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{z}_k - \mathbf{H}_k\mathbf{x}_k)$ denotes the whitened residual vector (see (14)).

Differentiating (49) yields the gradient

$$\nabla J_k(\mathbf{x}_k) = \mathbf{P}_{k|k-1}^{-1}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) - \frac{1}{\alpha}\mathbf{H}_k^\top \mathbf{R}_k^{-\frac{1}{2}} \boldsymbol{\psi}_k(\mathbf{x}_k), \tag{50}$$

where $\boldsymbol{\psi}_k(\mathbf{x}_k) \triangleq \tanh\left(\alpha\, \tilde{\mathbf{r}}_k(\mathbf{x}_k)\right)$ (applied componentwise), and the Hessian

$$\nabla^2 J_k(\mathbf{x}_k) = \mathbf{P}_{k|k-1}^{-1} + \mathbf{H}_k^\top \mathbf{R}_k^{-\frac{1}{2}} \mathbf{W}_k(\mathbf{x}_k) \mathbf{R}_k^{-\frac{1}{2}} \mathbf{H}_k, \tag{51}$$

where $\mathbf{W}_k(\mathbf{x}_k) \triangleq \operatorname{diag}\left(\operatorname{sech}^2\left(\alpha\, \tilde{\mathbf{r}}_k(\mathbf{x}_k)\right)\right)$ with $\operatorname{sech}^2(\cdot) = 1 - \tanh^2(\cdot)$ applied elementwise.

The explicit forms of the gradient and Hessian confirm that $J_k(\mathbf{x}_k)$ is twice continuously differentiable in a neighborhood of the minimizer $\hat{\mathbf{x}}_{k|k} = \arg\min_{\mathbf{x}_k} J_k(\mathbf{x}_k)$ (see (18)).

By interpreting the loss function $J_k(\mathbf{x}_k)$ as the negative log-posterior (up to an additive constant; see Section III-B), the posterior distribution can be written as

$$p(\mathbf{x}_k \mid \mathbf{z}_k) \propto \exp\left(-J_k(\mathbf{x}_k)\right), \tag{52}$$

where the normalizing constant is $Z_k \triangleq \int_{\mathbb{R}^n} \exp\left(-J_k(\mathbf{x})\right) \mathrm{d}\mathbf{x}$.

Given this twice differentiability, we apply the Laplace method [39] to approximate the posterior via a second-order

Taylor expansion of J_k around $\hat{\mathbf{x}}_{k|k}$, yielding a Gaussian density:

$$J_k(\mathbf{x}_k) = J_k(\hat{\mathbf{x}}_{k|k}) + \frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^\top \nabla^2 J_k(\hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}) + R(\mathbf{x}_k), \tag{53}$$

with $|R(\mathbf{x}_k)| = O\left(\|\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}\|^3\right)$.

To justify dropping the remainder term $R(\mathbf{x}_k)$, we examine the curvature of $J_k(\mathbf{x}_k)$. The prior term in (51) ensures strong convexity, as $\mathbf{P}_{k|k-1}^{-1} \succeq \lambda_{\min}\mathbf{I}$ with $\lambda_{\min} = 1/\lambda_{\max}(\mathbf{P}_{k|k-1}) > 0$, while the measurement term is positive semidefinite with entries bounded above by 1. As a result, the Hessian is uniformly lower-bounded: $\nabla^2 J_k(\mathbf{x}_k) \succeq \lambda_{\min}\mathbf{I}$. This guarantees that the posterior $p(\mathbf{x}_k \mid \mathbf{z}_k)$ is strongly log-concave and satisfies the concentration inequality

$$\mathbb{P}\left(\|\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}\| \geq r\right) \leq \exp\left(-\tfrac{1}{2}\lambda_{\min} r^2\right),$$

which implies that the posterior mass is sharply concentrated near $\hat{\mathbf{x}}_{k|k}$. Therefore, the contribution of the higher-order term $R(\mathbf{x}_k)$ in (53) becomes negligible.

Substituting the expansion (53) into (52) thus yields

$$p(\mathbf{x}_k \mid \mathbf{z}_k) \propto \exp\Big\{-J_k(\hat{\mathbf{x}}_{k|k}) - \frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^\top \nabla^2 J_k(\hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})\Big\}. \tag{54}$$

Since $J_k(\hat{\mathbf{x}}_{k|k})$ is constant in $\mathbf{x}_k$, it can be absorbed into the normalization. The resulting expression is the kernel of a Gaussian density with mean $\hat{\mathbf{x}}_{k|k}$ and covariance $\left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1}$, yielding the approximation

$$p(\mathbf{x}_k \mid \mathbf{z}_k) \approx \mathcal{N}\left(\hat{\mathbf{x}}_{k|k}, \left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1}\right). \tag{55}$$

The posterior error covariance is therefore approximated as

$$\mathbf{P}_{k|k} \approx \left[\nabla^2 J_k(\hat{\mathbf{x}}_{k|k})\right]^{-1}. \tag{56}$$

Appendix B
Proof of Theorem 1

To derive a regime for the robustness parameter α that ensures the stability of the robust filter under the conservative intermittent observation model, it is necessary to guarantee that the inlier probability λ_G, computed under a Gaussian approximation, exceeds the critical threshold $\overline{\lambda}_c$ established in [45] by a safety margin $\Delta > 0$, i.e., $\lambda_G \geq \overline{\lambda}_c + \Delta$. The inlier probability λ_G under the Gaussian model is given by

$$\lambda_G = \mathbb{P}\left(\|\tilde{\mathbf{r}}_k\|_\infty < \delta\right), \tag{57}$$

where $\delta = \frac{1}{\alpha}\operatorname{arccosh}\left(\theta^{-1/2}\right)$ is defined in (42) (with θ from (41)). Under the Gaussian surrogate model, the whitened residual satisfies $\tilde{\mathbf{r}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_r)$, as shown in the proof of Proposition 3. From the same result, a conservative lower bound on the inlier probability is given by

$$\lambda_G \geq \left[2\Phi\left(\frac{\delta}{\sigma_\rho}\right) - 1\right]^m, \tag{58}$$

where $\sigma_\rho = \sqrt{\rho(\mathbf{\Sigma}_r)}$ is defined in (43).

To ensure stability under this approximation, we require

$$\left[2\Phi\left(\frac{\delta}{\sigma_\rho}\right) - 1\right]^m \geq \overline{\lambda}_c + \Delta, \tag{59}$$

where $\Delta > 0$ compensates for the fact that heavy-tailed measurement noise may cause the true inlier probability λ_{HT} to fall below the Gaussian estimate λ_G. In practice, Δ may be selected empirically, as discussed in Section V-C. Taking the m-th root of both sides yields

$$2\Phi\left(\frac{\delta}{\sigma_\rho}\right) - 1 \geq (\overline{\lambda}_c + \Delta)^{1/m}. \tag{60}$$

Solving for the argument of $\Phi(\cdot)$, we obtain

$$\frac{\delta}{\sigma_\rho} \geq \Phi^{-1}\left(\frac{(\overline{\lambda}_c + \Delta)^{1/m} + 1}{2}\right). \tag{61}$$

Substituting the expression for δ into (61) and solving for α yields the final bound:

$$\alpha \leq \frac{\operatorname{arccosh}\left(\theta^{-1/2}\right)}{\sigma_\rho \cdot \Phi^{-1}\left(\frac{(\overline{\lambda}_c+\Delta)^{1/m}+1}{2}\right)}. \tag{62}$$

This upper bound ensures that the inlier probability meets the conservative stability condition required for mean-square boundedness of the robust filter, as stated in Theorem 1.

Chapter 5

Conclusion

This thesis presents the development and validation of a robust, fully distributed state estimation framework designed to function effectively in the presence of outliers, which are modeled by heavy-tailed noise distributions. The core of the proposed method is the integration of a smooth and convex log-cosh loss function within the generalized Bayesian inference framework. This approach successfully mitigates the impact of large residuals on the estimation process, which is a common problem in traditional Kalman Filter (KF). The resulting estimator preserves the recursive structure of the KF while effectively handling unreliable measurements using a bounded influence function.

A key theoretical contribution is the conservative stability analysis, which models the robust estimator as a KF with intermittent observations. This analysis led to an analytically derived upper bound for the robustness parameter, ensuring that the estimation error covariance remains bounded. For practical implementation in a distributed sensor network, the estimator leverages the Exact First Order Algorithm (EXTRA) algorithm for optimization and consensus. This allows for exact convergence to the optimal solution through local message exchanges and a fixed step size, making it computationally efficient.

Numerical evaluations conducted in a simulated target-tracking scenario consistently demonstrated the superior performance of the proposed method compared to both the standard KF and another robust estimator based on the Inverse Multi-Quadratic (IMQ) function. The proposed method achieved a reduction in the Root Mean Square Error (RMSE) of up to 20% when subjected to heavy-tailed Student t-distributed measurement noise and external disturbances. Furthermore, it requires fewer consensus iterations to converge, highlighting its communication efficiency. Overall, this thesis presents a resilient state estimation framework that is theoretically sound, computationally efficient, and empirically validated, offering a practical solution for modern distributed sensor networks operating in challenging and adversarial environments.

5-1 Limitations and Future Work

Although the proposed framework demonstrates some advancements in robust state estimation, certain limitations were identified, which suggest promising directions for future research.

A dynamic adjustment rule for the robustness parameter was introduced to enhance the adaptability. However, the empirical results from the simulations did not show a clear advantage of the adaptive approach over the static parameter. This suggests that although the concept is promising, its implementation may not be optimal. Further investigation into more sophisticated adaptive mechanisms is therefore warranted. Utilizing techniques for online hyperparameter tuning can fully realize the potential of an adaptive method.

Another limitation is the computational trade-off. In the centralized case, the proposed method replaces the linear update of the standard KF with an iterative optimization process. This is necessary for robustness but adds computational overhead that, while negligible for low-dimensional systems, can become significant in higher dimensions, making the filter slower than that of the standard KF.

The proposed framework was evaluated using static network topologies and time-invariant noise models. In many real-world applications, sensor networks are dynamic, with nodes moving, failing, or new nodes being added. The noise characteristics may also change over time. Future work could extend this framework to handle such dynamic scenarios, thereby broadening its applicability.

Furthermore, the evaluation was conducted under the assumption that every sensor node observed the same state components. A valuable direction for future work would be to assess the framework's performance with partial observability, where nodes measure only a subset of state variables or in networks with some "naive" nodes having no direct observations. Validating the algorithm under these challenging sensing conditions would be a significant extension.

Bibliography

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Commun. Mag.*, 40:102–114, Aug. 2002.

[2] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy, and M. S. Shehata. Structural health monitoring using wireless sensor networks: a comprehensive survey. *IEEE Commun. Surv. Tutor.*, 19:1403–1423, 2017.

[3] J. Sijs, M. Lazar, P. P. J. v. d. Bosch, and Z. Papp. An overview of non-centralized Kalman filters. In *Proc. IEEE Int. Conf. Control Appl.*, pages 738–744. IEEE, Jul. 2008.

[4] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. In *Proc. Ieee*, volume 95, pages 215–233. IEEE, Jan. 2007.

[5] R. Olfati-Saber. Distributed Kalman filtering for sensor networks. In *Proc. IEEE Conf. Decis. Control*, pages 5492–5498. IEEE, 2007.

[6] F. S. Cattivelli and A. H. Sayed. Diffusion strategies for distributed Kalman filtering and smoothing. *IEEE Trans. Autom. Control*, 55:2069–2084, Sep. 2010.

[7] A. G. Dimakis, A. D. Sarwate, and M. J. Wainwright. Geographic gossip: efficient aggregation for sensor networks. In *Proc. Int. Conf. Inf. Process. Sens. Netw. (IPSN)*, pages 69–76, New York, NY, 2006. ACM.

[8] V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Chichester [u.a.], 3rd edition, 1994.

[9] H. Zhu, M. J. V. Amuri, X. Li, and J. Shen. Mean-shift-based outliers-robust distributed Kalman filter for wireless sensor network systems. *IEEE Trans. Instrum. Meas.*, 74:1–12, 2025.

[10] R. Olfati-Saber. Distributed Kalman filter with embedded consensus filters. In *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, volume 2005, pages 8179–8184, [Piscataway, NJ], 2005. IEEE.

[11] A. T. Kamal, C. Ding, B. Song, J. A. Farrell, and A. K. Roy-Chowdhury. A generalized Kalman consensus filter for wide-area video networks. In *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, pages 7863–7869. IEEE, Dec. 2011.

[12] R. Olfati-Saber and J. S. Shamma. Consensus filters for sensor networks and distributed sensor fusion. In *Proc. IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, volume 2005, pages 6698–6703, [Piscataway, NJ], 2005. IEEE.

[13] P. Dong, Z. Jing, H. Leung, K. Shen, and M. Li. Robust consensus nonlinear information filter for distributed sensor networks with measurement outliers. *IEEE Trans. Cybern.*, 49:3731–3743, Oct. 2019.

[14] S. Modalavalsa, U. K. Sahoo, A. K. Sahoo, and S. Kumar. Diffusion minimum generalized rank norm over distributed adaptive networks: formulation and performance analysis. *IEEE Trans. Signal Inf. Process. Netw.*, 5:669–683, Dec. 2019.

[15] G. Duran-Martin et al. Outlier-robust Kalman filtering through generalised Bayes. In R. Salakhutdinov et al., editors, *Proc. Int. Conf. Mach. Learn.*, volume 235, pages 12138–12171. PMLR, Jul. 2024.

[16] W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. arXiv preprint arXiv:1404.6264, 2014.

[17] L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pages 63–70, 2005.

[18] M. Fiedler. Algebraic connectivity of graphs. *Czechoslov. Math. J.*, 23:298–305, 1973.

[19] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems Control Letters*, 53(1):65–78, 2004.

[20] S. J. Julier and J. K. Uhlmann. A non-divergent estimation algorithm in the presence of unknown correlations. In *Proc. Am. Control Conf.*, volume 4, pages 2369–2373. IEEE, 1997.

[21] G. Battistelli, L. Chisci, G. Mugnai, A. Farina, and A. Graziano. Consensus-based linear and nonlinear filtering. *IEEE Trans. Autom. Control*, 60:1410–1415, May 2015.

[22] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.*, 81:82–86, Mar. 1986.

[23] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, March 1964.

[24] K. Gokcesu and H. Gokcesu. Generalized huber loss for robust learning and its efficient minimization for a robust statistics. Aug. 2021.

[25] Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, 1977.

[26] Albert E. Beaton and John W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[27] M. O. Sayin, N. D. Vanli, and S. S. Kozat. A novel family of adaptive filtering algorithms based on the logarithmic cost. *IEEE Trans. Signal Process.*, 62:4411–4424, Sep. 2014.

[28] Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, and Johan A.K. Suykens. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16(30):993–1034, 2015.

[29] B. Chen, X. Liu, H. Zhao, and J. C. Príncipe. Maximum correntropy kalman filter. arXiv preprint arXiv:1509.04580, 2015.

[30] B. Chen, L. Xing, B. Xu, H. Zhao, N. Zheng, and J. C. Principe. Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering. *IEEE Trans. Signal Process.*, 65:2888–2901, Jul. 2017.

[31] Chien-Hao Tseng, Sheng-Fuu Lin, and Dah-Jing Jwo. Robust huber-based cubature kalman filter for gps navigation processing. *Journal of Navigation*, 70(3):527–546, 2017.

[32] Fuzhi Zhang, Shuangxia Sun, and Huawei Yi. Robust collaborative recommendation algorithm based on kernel function and welsch reweighted m-estimator. *IET Information Security*, 9:257–265, 2015.

[33] Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab. Robust optimization for deep regression. ICCV '15, page 2830–2838, USA, 2015. IEEE Computer Society.

[34] Weifeng Liu, Puskal P. Pokharel, and Jose C. Principe. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing*, 55(11):5286–5298, 2007.

[35] Jun Qi, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, 27:1485–1489, 2020.

[36] C. Liu and M. Jiang. Robust adaptive filter with lncosh cost. *Signal Process.*, 168:107348, Mar. 2020.

[37] S. Wang, W. Wang, K. Xiong, H. H. C. Iu, and C. K. Tse. Logarithmic hyperbolic cosine adaptive filter and its performance analysis. *IEEE Trans. Syst., Man, Cybern., Syst.*, 51:2512–2524, Apr. 2021.

[38] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 02 2016.

[39] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry. Kalman filtering with intermittent observations. *IEEE Trans. Autom. Control*, 49:1453–1464, Sep. 2004.

Glossary

List of Acronyms

WSN	Wireless Sensor Network
FC	Fusion Center
KF	Kalman Filter
CKF	Centralized Kalman Filter
RMSE	Root Mean Square Error
MMSE	Minimum Mean Squared Estimate
MSE	Mean Squared Error
MAE	Mean Absolute Error
LLAD	Least Logarithmic Absolute Difference
MCC	Maximum Correntropy Criterion
KRSL	Kernel Risk-Sensitive Loss
LHCAF	Logarithmic Hyperbolic Cosine Adaptive Filter
GKCF	Generalized Kalman Consensus Filter
DKF	Distributed Kalman Filter
MAP	Maximum A Posteriori
HCMCI	Hybrid Consensus on Measurement and Information
MSD	Mean Square Deviation
RCNIF	Robust Consensus Nonlinear Information Filter
DoF	Degree of Freedom
EM	Expectation-Maximization
VB	Variational Bayesian
KL	Kullback-Leibler
ELBO	Evidence Lower Bound
dMGRN	Diffusion Minimum Generalized Rank Norm

MVE	Minimum Volume Ellipsoid
SSMs	State Space Models
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
WoLF	Weighted Observation Likelihood Filter
PIF	Posterior Influence Function
EXTRA	Exact First Order Algorithm
IMQ	Inverse Multi-Quadratic
MD	Mahalanobis Distance
TMD	Threshold Mahalanobis Distance
LLA	Least log-cosh Algorithm
EMSE	Excess Mean Square Error