Decentralized Data Fusion and Active Sensing with Mobile Sensors for Modeling and Predicting Spatiotemporal Traffic Phenomena

Jie Chen[†], **Kian Hsiang Low**[†], **Colin Keng-Yan Tan**[†], **Ali Oran**[§]
Dept. Computer Science, National University of Singapore[†]
Singapore-MIT Alliance for Research and Technology[§]
Republic of Singapore

Patrick Jaillet[‡], **John Dolan**[¶], **Gaurav Sukhatme**[*]
Dept. Electrical Engineering and Computer Science, MIT[‡]
The Robotics Institute, Carnegie Mellon University[¶]
Dept. Computer Science, University of Southern California[*]
USA

Abstract

The problem of modeling and predicting spatiotemporal traffic phenomena over an urban road network is important to many traffic applications such as detecting and forecasting congestion hotspots. This paper presents a decentralized data fusion and active sensing (D^{2}FAS) algorithm for mobile sensors to actively explore the road network to gather and assimilate the most informative data for predicting the traffic phenomenon. We analyze the time and communication complexity of D^{2}FAS and demonstrate that it can scale well with a large number of observations and sensors. We provide a theoretical guarantee on its predictive performance to be equivalent to that of a sophisticated centralized sparse approximation for the Gaussian process (GP) model: The computation of such a sparse approximate GP model can thus be parallelized and distributed among the mobile sensors (in a Google-like MapReduce paradigm), thereby achieving efficient and scalable prediction. We also theoretically guarantee its active sensing performance that improves under various practical environmental conditions. Empirical evaluation on real-world urban road network data shows that our D^{2}FAS algorithm is significantly more time-efficient and scalable than state-of-the-art centralized algorithms while achieving comparable predictive performance.

1 Introduction

Knowing and understanding the traffic conditions and phenomena over road networks has become increasingly important to the goal of achieving smooth-flowing, congestion-free traffic, especially in densely-populated urban cities. According to a 2011 urban mobility report (Schrank *et al.*, 2011), traffic congestion in the USA has caused 1.9 billion gallons of extra fuel, 4.8 billion hours of travel delay, and \$101 billion of delay and fuel cost. Such huge resource wastage can potentially be mitigated if the spatiotemporally varying traffic phenomena (e.g., speeds and travel times along road segments) are predicted accurately enough in real time to detect and forecast the congestion hotspots; network-level (e.g., ramp metering, road pricing) and user-level (e.g., route replanning) measures can then be taken to relieve this congestion, so as to improve the overall efficiency of road networks.

In practice, it is non-trivial to achieve real-time, accurate prediction of a spatiotemporally varying traffic phenomenon because the quantity of sensors that can be deployed to observe an entire road network is cost-constrained. Traditionally, static sensors such as loop detectors (Krause *et al.*, 2008a; Wang and Papageorgiou, 2005) are placed at designated locations in a road network to collect data for predicting the traffic phenomenon. However, they provide sparse coverage (i.e., many road segments are not observed, thus leading to data sparsity), incur high installation and maintenance costs, and cannot reposition by themselves in response to changes in the traffic phenomenon. Low-cost GPS technology allows the collection of traffic data using passive mobile probes (Work *et al.*, 2010) (e.g., taxis/cabs). Unlike static sensors, they can directly measure the travel times along road segments. But, they provide fairly sparse coverage due to low GPS sampling frequency (i.e., often imposed by taxi/cab companies) and no control over their routes, incur high initial implementation cost, pose privacy issues, and produce highly-varying speeds and travel times while traversing the same road segment due to inconsistent driving behaviors. A critical mass of probes is needed on each road segment to ease the severity of the last drawback (Srinivasan and Jovanis, 1996) but is often hard to achieve on non-highway segments due to sparse coverage. In contrast, we propose the use of active mobile probes (Turner *et al.*, 1998) to overcome the limitations of static and passive mobile probes. In particular, they can be directed to explore any segments of a road network to gather traffic data at a desired GPS sampling rate while enforcing consistent driving behavior.

How then do the mobile probes/sensors actively explore a road network to gather and assimilate the most informative observations for predicting the traffic phenomenon? There are three key issues surrounding this problem, which will be discussed together with the related works:

Models for predicting spatiotemporal traffic phenomena. The spatiotemporal correlation structure of a traffic phenomenon can be exploited to predict the traffic conditions of any unobserved road segment at any time using the observations taken along the sensors' paths. To achieve this, existing Bayesian filtering frameworks (Chen *et al.*, 2011; Wang and Papageorgiou, 2005; Work *et al.*, 2010) utilize various handcrafted parametric models predicting traffic flow along a highway stretch that only correlate adjacent segments of the highway. So, their predictive performance will be compromised when the current observations are sparse and/or the actual spatial correlation spans multiple segments. Their strong Markov assumption further exacerbates this problem. It is also not shown how these models can be generalized to work for arbitrary road network topologies and more complex correlation structure. Existing multivariate parametric traffic prediction models (Kamarianakis and Prastacos, 2003; Min and Wynter, 2011) do not quantify uncertainty estimates of the predictions and impose rigid spatial locality assumptions that do not adapt to the true underlying correlation structure.

In contrast, we assume the traffic phenomenon over an urban road network (i.e., comprising the full range of road types like highways, arterials, slip roads) to be realized from a rich class of Bayesian non-parametric models called the *Gaussian process* (GP) (Section 2) that can formally characterize its spatiotemporal correlation structure and be refined with growing number of observations. More importantly, GP can provide formal measures of predictive uncertainty (e.g., based on variance or entropy criterion) for directing the sensors to explore highly uncertain areas of the road network. Krause *et al.* (2008a) used GP to represent the traffic phenomenon over a network of only highways and defined the correlation of speeds between highway segments to depend only on the geodesic (i.e., shortest path) distance of these segments with respect to the network topology; their features are not considered. Neumann *et al.* (2009) maintained a mixture of two independent GPs for flow prediction such that the correlation structure of one GP utilized road segment features while that of the other GP depended on manually specified relations (instead of geodesic distance) between segments with respect to an undirected network topology. Different from the above works, we propose a relational GP whose correlation structure exploits the geodesic distance between segments based on the topology of a directed road network with vertices denoting road segments and edges indicating adjacent segments weighted by dissimilarity of their features, hence tightly integrating the features and relational information.

Data fusion. The observations are gathered distributedly by each sensor along its path in the road network and have to be assimilated in order to predict the traffic phenomenon. Since a large number of observations are expected to be collected, a centralized approach to GP prediction cannot be performed in real time due to its cubic time complexity.

To resolve this, we propose a decentralized data fusion approach to efficient and scalable approximate GP prediction (Section 3). Existing decentralized and distributed Bayesian filtering frameworks for addressing non-traffic related problems (Chung *et al.*, 2004; Coates, 2004; Olfati-Saber, 2005; Rosencrantz *et al.*, 2003; Sukkarieh *et al.*, 2003) will face the same difficulties as their centralized counterparts described above if applied to predicting traffic phenomena, thus resulting in loss of predictive performance. Distributed regression algorithms (Guestrin *et al.*, 2004; Paskin and Guestrin, 2004) for static sensor networks gain efficiency from spatial locality assumptions, which cannot be exploited by mobile sensors whose paths are not constrained by locality. Cortes (2009) proposed a distributed data fusion approach to approximate GP prediction based on an iterative Jacobi overrelaxation algorithm, which incurs some critical limitations: (a) the past observations taken along the sensors' paths are assumed to be uncorrelated, which greatly undermines its predictive performance when they are in fact correlated and/or the current observations are sparse; (b) when the number of sensors grows large, it converges very slowly; (c) it assumes that the range of positive correlation has to be bounded by some factor of the communication range. Our proposed decentralized algorithm does not suffer from these limitations and can be computed exactly with efficient time bounds.

Active sensing. The sensors have to coordinate to actively gather the most informative observations for minimizing the uncertainty of modeling and predicting the traffic phenomenon. Existing centralized (Low *et al.*, 2008, 2009a, 2011) and decentralized (Low *et al.*, 2012; Stranders *et al.*, 2009) active sensing algorithms scale poorly with a large number of observations and sensors. We propose a partially decentralized active sensing algorithm that overcomes these issues of scalability (Section 4).

This paper presents a novel *Decentralized Data Fusion and Active Sensing* (D^{2}FAS) algorithm (Sections 3 and 4) for sampling spatiotemporally varying environmental phenomena with mobile sensors. Note that the decentralized data fusion component of D^{2}FAS can also be used for static and passive mobile sensors. The practical applicability of D^{2}FAS is not restricted to traffic monitoring; it can be used in other environmental sensing applications such as mineral prospecting (Low *et al.*, 2007), monitoring of ocean and freshwater phenomena (Dolan *et al.*, 2009; Podnar *et al.*, 2010; Low *et al.*, 2009b, 2011, 2012) (e.g., plankton bloom, anoxic zones), forest ecosystems, pollution (e.g., oil spill), or contamination (e.g., radiation leak). The specific contributions of this paper include:

- Analyzing the time and communication overheads of

D^{2}FAS (Section 5): We prove that D^{2}FAS can scale better than existing state-of-the-art centralized algorithms with a large number of observations and sensors;

- Theoretically guaranteeing the predictive performance of the decentralized data fusion component of D^{2}FAS to be equivalent to that of a sophisticated centralized sparse approximation for the GP model (Section 3): The computation of such a sparse approximate GP model can thus be parallelized and distributed among the mobile sensors (in a Google-like MapReduce paradigm), thereby achieving efficient and scalable prediction;
- Theoretically guaranteeing the performance of the partially decentralized active sensing component of D^{2}FAS, from which various practical environmental conditions can be established to improve its performance;
- Developing a relational GP model whose correlation structure can exploit both the road segment features and road network topology information (Section 2.1);
- Empirically evaluating the predictive performance, time efficiency, and scalability of the D^{2}FAS algorithm on a real-world traffic phenomenon (i.e., speeds of road segments) dataset over an urban road network (Section 6): D^{2}FAS is more time-efficient and scales significantly better with increasing number of observations and sensors while achieving predictive performance close to that of existing state-of-the-art centralized algorithms.

2 Relational Gaussian Process Regression

The *Gaussian process* (GP) can be used to model a spatiotemporal traffic phenomenon over a road network as follows: The traffic phenomenon is defined to vary as a realization of a GP. Let V be a set of road segments representing the domain of the road network such that each road segment $s \in V$ is specified by a p-dimensional vector of features and is associated with a realized (random) measurement z_s (Z_s) of the traffic condition such as speed if s is observed (unobserved). Let $\{Z_s\}_{s\in V}$ denote a GP, that is, every finite subset of $\{Z_s\}_{s\in V}$ follows a multivariate Gaussian distribution (Rasmussen and Williams, 2006). Then, the GP is fully specified by its *prior* mean $\mu_s \triangleq \mathbb{E}[Z_s]$ and covariance $\sigma_{ss'} \triangleq \text{cov}[Z_s, Z_{s'}]$ for all $s, s' \in V$. In particular, we will describe in Section 2.1 how the covariance $\sigma_{ss'}$ for modeling the correlation of measurements between all pairs of segments $s, s' \in V$ can be designed to exploit the road segment features and the road network topology.

A chief capability of the GP model is that of performing probabilistic regression: Given a set $D \subset V$ of observed road segments and a column vector z_D of corresponding measurements, the joint distribution of the measurements at any set $Y \subseteq V \setminus D$ of unobserved road segments remains Gaussian with the following *posterior* mean vector and covariance matrix

$$\mu_{Y|D} \triangleq \mu_Y + \Sigma_{YD}\Sigma_{DD}^{-1}(z_D - \mu_D) \qquad (1)$$

$$\Sigma_{YY|D} \triangleq \Sigma_{YY} - \Sigma_{YD}\Sigma_{DD}^{-1}\Sigma_{DY} \qquad (2)$$

where μ_Y (μ_D) is a column vector with mean components μ_s for all $s \in Y$ ($s \in D$), Σ_{YD} (Σ_{DD}) is a covariance matrix with covariance components $\sigma_{ss'}$ for all $s \in Y, s' \in D$ ($s, s' \in D$), and Σ_{DY} is the transpose of Σ_{YD}. The posterior mean vector $\mu_{Y|D}$ (1) is used to predict the measurements at any set Y of unobserved road segments. The posterior covariance matrix $\Sigma_{YY|D}$ (2), which is independent of the measurements z_D, can be processed in two ways to quantify the uncertainty of these predictions: (a) the trace of $\Sigma_{YY|D}$ yields the sum of posterior variances $\Sigma_{ss|D}$ over all $s \in Y$; (b) the determinant of $\Sigma_{YY|D}$ is used in calculating the Gaussian posterior joint entropy

$$\mathbb{H}[Z_Y|Z_D] \triangleq \frac{1}{2}\log(2\pi e)^{|Y|}\left|\Sigma_{YY|D}\right| . \qquad (3)$$

In contrast to the first measure of uncertainty that assumes conditional independence between measurements in the set Y of unobserved road segments, the entropy-based measure (3) accounts for their correlation, thereby not overestimating their uncertainty. Hence, we will focus on using the entropy-based measure of uncertainty in this paper.

2.1 Graph-Based Kernel

If the observations are noisy (i.e., by assuming additive independent identically distributed Gaussian noise with variance σ_n^2), then their prior covariance $\sigma_{ss'}$ can be expressed as $\sigma_{ss'} = k(s, s') + \sigma_n^2\delta_{ss'}$ where $\delta_{ss'}$ is a Kronecker delta that is 1 if $s = s'$ and 0 otherwise, and k is a kernel function measuring the pairwise "similarity" of road segments. For a traffic phenomenon (e.g., road speeds), the correlation of measurements between pairs of road segments depends not only on their features (e.g., length, number of lanes, speed limit, direction) but also the road network topology. So, the kernel function is defined to exploit both the features and topology information, which will be described next.

Definition 1 (Road Network) *Let the road network be represented as a weighted directed graph $G \triangleq (V, E, m)$ that consists of*

- *a set V of vertices denoting the domain of all possible road segments,*
- *a set $E \subseteq V \times V$ of edges where there is an edge (s, s') from $s \in V$ to $s' \in V$ iff the end of segment s connects to the start of segment s' in the road network, and*
- *a weight function $m : E \to \mathbb{R}^+$ measuring the standardized Manhattan distance (Borg and Groenen, 2005) $m((s, s')) \triangleq \sum_{i=1}^{p} |[s]_i - [s']_i|/r_i$ of each edge (s, s') where $[s]_i$ ($[s']_i$) is the i-th component of the feature vector specifying road segment s (s'), and r_i is the range of the i-th feature. The weight function m serves as a dissimilarity measure between adjacent road segments.*

The next step is to compute the shortest path distance $d(s, s')$ between all pairs of road segments $s, s' \in V$ (i.e., using Floyd-Warshall or Johnson's algorithm) with respect to the topology of the weighted directed graph G. Such a distance function is again a measure of dissimilarity, rather than one of similarity, as required by a kernel function. Furthermore, a valid GP kernel needs to be positive semidefinite and symmetric (Schölkopf and Smola, 2002), which are clearly violated by d.

To construct a valid GP kernel from d, multi-dimensional scaling (Borg and Groenen, 2005) is applied to embed the domain of road segments into the p'-dimensional Euclidean space $\mathbb{R}^{p'}$. Specifically, a mapping $g : V \rightarrow \mathbb{R}^{p'}$ is determined by minimizing the squared loss $g^* = \arg\min_g \sum_{s,s' \in V}(d(s,s') - \|g(s) - g(s')\|)^2$. With a small squared loss, the Euclidean distance $\|g^*(s) - g^*(s')\|$ between $g^*(s)$ and $g^*(s')$ is expected to closely approximate the shortest path distance $d(s,s')$ between any pair of road segments s and s'. After embedding into Euclidean space, a conventional kernel function such as the squared exponential one (Rasmussen and Williams, 2006) can be used:

$$k(s,s') = \sigma_s^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{p'}\left(\frac{[g^*(s)]_i - [g^*(s')]_i}{\ell_i}\right)^2\right)$$

where $[g^*(s)]_i$ ($[g^*(s')]_i$) is the i-th component of the p'-dimensional vector $g^*(s)$ ($g^*(s')$), and the hyperparameters $\sigma_s, \ell_1, \ldots, \ell_{p'}$ are, respectively, signal variance and length-scales that can be learned using maximum likelihood estimation (Rasmussen and Williams, 2006). The resulting kernel function k[1] is guaranteed to be valid.

2.2 Subset of Data Approximation

Although the GP is an effective predictive model, it faces a practical limitation of cubic time complexity in the number $|D|$ of observations; this can be observed from computing the posterior distribution (i.e., (1) and (2)), which requires inverting covariance matrix Σ_{DD} and incurs $\mathcal{O}(|D|^3)$ time. If $|D|$ is expected to be large, GP prediction cannot be performed in real time. For practical usage, we have to resort to computationally cheaper approximate GP prediction.

A simple method of approximation is to select only a subset U of the entire set D of observed road segments (i.e., $U \subset D$) to compute the posterior distribution of the measurements at any set $Y \subseteq V \setminus D$ of unobserved road segments. Such a *subset of data* (SoD) approximation method produces the following predictive Gaussian distribution, which closely resembles that of the full GP model (i.e., by simply replacing D in (1) and (2) with U):

$$\mu_{Y|U} = \mu_Y + \Sigma_{YU}\Sigma_{UU}^{-1}(z_U - \mu_U) \quad (4)$$

$$\Sigma_{YY|U} = \Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY} \; . \quad (5)$$

Notice that the covariance matrix Σ_{UU} to be inverted only incurs $\mathcal{O}(|U|^3)$ time, which is independent of $|D|$.

The predictive performance of SoD approximation is sensitive to the selection of subset U. In practice, random subset selection often yields poor performance. This issue can be resolved by actively selecting an informative subset U in an iterative greedy manner: Firstly, U is initialized to be an empty set. Then, all road segments in $D \setminus U$ are scored based on a criterion that can be chosen from, for example, the works of (Krause *et al.*, 2008b; Lawrence *et al.*, 2003; Seeger and Williams, 2003). The highest-scored segment is selected for inclusion in U and removed from D. This greedy selection procedure is iterated until U reaches a predefined size. Among the various criteria introduced earlier, the differential entropy score (Lawrence *et al.*, 2003) is reported to perform well (Oh *et al.*, 2010); it is a monotonic function of the posterior variance $\Sigma_{ss|U}$ (5), thus resulting in the greedy selection of a segment $s \in D \setminus U$ with the largest variance in each iteration.

3 Decentralized Data Fusion

In the previous section, two centralized data fusion approaches to exact (i.e., (1) and (2)) and approximate (i.e., (4) and (5)) GP prediction are introduced. In this section, we will discuss the decentralized data fusion component of our D^{2}FAS algorithm, which distributes the computational load among the mobile sensors to achieve efficient and scalable approximate GP prediction.

The intuition of our decentralized data fusion algorithm is as follows: Each of the K mobile sensors constructs a local summary of the observations taken along its own path in the road network and communicates its local summary to every other sensor. Then, it assimilates the local summaries received from the other sensors into a globally consistent summary, which is exploited for predicting the traffic phenomenon as well as active sensing. This intuition will be formally realized and described in the paragraphs below.

While exploring the road network, each mobile sensor summarizes its local observations taken along its path based on a common support set $U \subset V$ known to all the other sensors. Its local summary is defined as follows:

Definition 2 (Local Summary) *Given a common support set $U \subset V$ known to all K mobile sensors, a set $D_k \subset V$ of observed road segments and a column vector z_{D_k} of corresponding measurements local to mobile sensor k, its local summary is defined as a tuple $(\dot{z}_U^k, \dot{\Sigma}_{UU}^k)$ where*

$$\dot{z}_U^k \triangleq \Sigma_{UD_k}\Sigma_{D_kD_k|U}^{-1}(z_{D_k} - \mu_{D_k}) \quad (6)$$

$$\dot{\Sigma}_{UU}^k \triangleq \Sigma_{UD_k}\Sigma_{D_kD_k|U}^{-1}\Sigma_{D_kU} \quad (7)$$

such that $\Sigma_{D_kD_k|U}$ is defined in a similar manner to (5).

Remark. Unlike SoD (Section 2.2), the support set U of road segments does not have to be observed, since the local summary (i.e., (6) and (7)) is independent of the corresponding measurements z_U. So, U does not need to be a subset of $D = \bigcup_{k=1}^K D_k$. To select an informative support set U from the set V of all possible segments in the road network, an offline active selection procedure similar to that in the last paragraph of Section 2.2 can be performed just once prior to observing data to determine U. In contrast, SoD has to perform online active selection every time new road segments are being observed.

[1] For spatiotemporal traffic modeling, the kernel function k can be extended to account for the temporal dimension.

By communicating its local summary to every other sensor, each mobile sensor can then construct a globally consistent summary from the received local summaries:

Definition 3 (Global Summary) *Given a common support set $U \subset V$ known to all K mobile sensors and the local summary $(\dot{z}_U^k, \dot{\Sigma}_{UU}^k)$ of every mobile sensor $k = 1, \ldots, K$, the global summary is defined as a tuple $(\ddot{z}_U, \ddot{\Sigma}_{UU})$ where*

$$\ddot{z}_U \triangleq \sum_{k=1}^{K} \dot{z}_U^k \tag{8}$$

$$\ddot{\Sigma}_{UU} \triangleq \Sigma_{UU} + \sum_{k=1}^{K} \dot{\Sigma}_{UU}^k \ . \tag{9}$$

Remark. In this paper, we assume all-to-all communication between the K mobile sensors. Supposing this is not possible and each sensor can only communicate locally with its neighbors, the summation structure of the global summary (specifically, (8) and (9)) makes it amenable to be constructed using distributed consensus filters (Olfati-Saber, 2005). We omit these details since they are beyond the scope of this paper.

Finally, the global summary is exploited by each mobile sensor to compute a globally consistent predictive Gaussian distribution, as detailed in Theorem 1A below, as well as to perform decentralized active sensing (Section 4):

Theorem 1 *Let a common support set $U \subset V$ be known to all K mobile sensors.*

A. *Given the global summary $(\ddot{z}_U, \ddot{\Sigma}_{UU})$, each mobile sensor computes a globally consistent predictive Gaussian distribution $\mathcal{N}(\overline{\mu}_Y, \overline{\Sigma}_{YY})$ of the measurements at any set Y of unobserved road segments where*

$$\overline{\mu}_Y \triangleq \mu_Y + \Sigma_{YU}\ddot{\Sigma}_{UU}^{-1}\ddot{z}_U \tag{10}$$

$$\overline{\Sigma}_{YY} \triangleq \Sigma_{YY} - \Sigma_{YU}(\Sigma_{UU}^{-1} - \ddot{\Sigma}_{UU}^{-1})\Sigma_{UY} \ . \tag{11}$$

B. *Let $\mathcal{N}(\mu_{Y|D}^{\text{PITC}}, \Sigma_{YY|D}^{\text{PITC}})$ be the predictive Gaussian distribution computed by the centralized sparse partially independent training conditional (PITC) approximation of GP model (Quiñonero-Candela and Rasmussen, 2005) where*

$$\mu_{Y|D}^{\text{PITC}} \triangleq \mu_Y + \Gamma_{YD}\left(\Gamma_{DD} + \Lambda\right)^{-1}(z_D - \mu_D) \tag{12}$$

$$\Sigma_{YY|D}^{\text{PITC}} \triangleq \Sigma_{YY} - \Gamma_{YD}\left(\Gamma_{DD} + \Lambda\right)^{-1}\Gamma_{DY} \tag{13}$$

such that

$$\Gamma_{BB'} \triangleq \Sigma_{BU}\Sigma_{UU}^{-1}\Sigma_{UB'} \tag{14}$$

and Λ is a block-diagonal matrix constructed from the K diagonal blocks of $\Sigma_{DD|U}$, each of which is a matrix $\Sigma_{D_kD_k|U}$ for $k = 1, \ldots, K$ where $D = \bigcup_{k=1}^K D_k$. Then, $\overline{\mu}_Y = \mu_{Y|D}^{\text{PITC}}$ and $\overline{\Sigma}_{YY} = \Sigma_{YY|D}^{\text{PITC}}$.

The proof of Theorem 1B is given in Appendix A. The equivalence result of Theorem 1B bears two implications:

Remark 1. The computation of PITC can be parallelized and distributed among the mobile sensors in a Google-like MapReduce paradigm (Chu *et al.*, 2007), thereby improving the time efficiency of prediction: Each of the K mappers (sensors) is tasked to compute its local summary while the reducer (any sensor) sums these local summaries into a global summary, which is then used to compute the predictive Gaussian distribution. Supposing $|Y| \leq |U|$ for simplicity, the $\mathcal{O}\big(|D|((|D|/K)^2 + |U|^2)\big)$ time incurred by PITC can be reduced to $\mathcal{O}\big((|D|/K)^3 + |U|^3 + |U|^2K\big)$ time of running our decentralized algorithm on each of the K sensors, the latter of which scales better with increasing number $|D|$ of observations.

Remark 2. We can draw insights from PITC to elucidate an underlying property of our decentralized algorithm: It is assumed that $Z_{D_1}, \ldots, Z_{D_K}, Z_Y$ are conditionally independent given the measurements at the support set U of road segments. To potentially reduce the degree of violation of this assumption, an informative support set U is actively selected, as described earlier in this section. Furthermore, the experimental results on real-world urban road network data[2] (Section 6) show that D^{2}FAS can achieve predictive performance comparable to that of the full GP model while enjoying significantly lower computational cost, thus demonstrating the practicality of such an assumption for predicting traffic phenomena. The predictive performance of D^{2}FAS can be improved by increasing the size of U at the expense of greater time and communication overhead.

4 Decentralized Active Sensing

The problem of active sensing with K mobile sensors is formulated as follows: Given the set $D_k \subset V$ of observed road segments and the currently traversed road segment $s_k \in V$ of every mobile sensor $k = 1, \ldots, K$, the mobile sensors have to coordinate to select the most informative walks $w_1^*, \ldots, w_K^*$ of length (i.e., number of road segments) L each and with respective origins $s_1, \ldots, s_K$ in the road network G:

$$(w_1^*, \ldots, w_K^*) = \mathop{\arg\max}_{(w_1,\ldots,w_K)} \mathbb{H}\left[Z_{\bigcup_{k=1}^K Y_{w_k}} \middle| Z_{\bigcup_{k=1}^K D_k}\right] \tag{15}$$

where Y_{w_k} denotes the set of unobserved road segments induced by the walk w_k. To simplify notation, let a joint walk be denoted by $w \triangleq (w_1, \ldots, w_K)$ (similarly, for w^*) and its induced set of unobserved road segments be $Y_w \triangleq \bigcup_{k=1}^K Y_{w_k}$ from now on. Interestingly, it can be shown using the chain rule for entropy that these maximum-entropy walks w^* minimize the posterior joint entropy (i.e., $\mathbb{H}[Z_{V \setminus (D \bigcup Y_{w^*})} | Z_{D \bigcup Y_{w^*}}]$) of the measurements at the remaining unobserved segments (i.e., $V \setminus (D \bigcup Y_{w^*})$) in the road network. After executing the walk w_k^*, each mobile sensor k observes the set $Y_{w_k^*}$ of road segments and updates its local information:

$$D_k \leftarrow D_k \bigcup Y_{w_k^*}\,, z_{D_k} \leftarrow z_{D_k \bigcup Y_{w_k^*}}\,, s_k \leftarrow \text{terminus of } w_k^* \ . \tag{16}$$

[2] Quiñonero-Candela and Rasmussen (2005) only illustrated the predictive performance of PITC on a simulated toy example.

Evaluating the Gaussian posterior entropy term in (15) involves computing a posterior covariance matrix (3) using one of the data fusion methods described earlier: If (2) of full GP model (Section 2) or (5) of SoD (Section 2.2) is to be used, then the observations that are gathered distributedly by the sensors have to be fully communicated to a central data fusion center. In contrast, our decentralized data fusion algorithm (Section 3) only requires communicating local summaries (Definition 2) to compute (11) for solving the active sensing problem (15):

$$w^* = \arg\max_{w} \overline{\mathbb{H}}[Z_{Y_w}] \ , \tag{17}$$

$$\overline{\mathbb{H}}[Z_{Y_w}] \triangleq \frac{1}{2}\log(2\pi e)^{|Y_w|}\left|\overline{\Sigma}_{Y_wY_w}\right| \ . \tag{18}$$

Without imposing any structural assumption, solving the active sensing problem (17) will be prohibitively expensive due to the space of possible joint walks w that grows exponentially in the number K of mobile sensors. To overcome this scalability issue for D^{2}FAS, our key idea is to construct a block-diagonal matrix whose log-determinant closely approximates that of $\overline{\Sigma}_{Y_wY_w}$ (11) and exploit the property that the log-determinant of such a block-diagonal matrix can be decomposed into a sum of log-determinants of its diagonal blocks, each of which depends only on the walks of a disjoint subset of the K mobile sensors. Consequently, the active sensing problem can be partially decentralized, leading to a reduced space of possible joint walks to be searched, as detailed in the rest of this section.

Firstly, we extend an earlier assumption in Section 3: $Z_{D_1},\ldots,Z_{D_K},Z_{Y_{w_1}},\ldots,Z_{Y_{w_K}}$ are conditionally independent given the measurements at the support set U of road segments. Then, it can be shown via the equivalence to PITC (Theorem 1B) that $\overline{\Sigma}_{Y_wY_w}$ (11) comprises diagonal blocks of the form $\overline{\Sigma}_{Y_{w_k}Y_{w_k}}$ for $k=1,\ldots,K$ and off-diagonal blocks of the form $\Sigma_{Y_{w_k}U}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UY_{w_{k'}}}$ for $k,k'=1,\ldots,K$ and $k\neq k'$. In particular, each off-diagonal block of $\overline{\Sigma}_{Y_wY_w}$ represents the correlation of measurements between the unobserved road segments Y_{w_k} and $Y_{w_{k'}}$ along the respective walks w_k of sensor k and $w_{k'}$ of sensor k'. If the correlation between some pair of their possible walks is high enough, then their walks have to be coordinated. This is formally realized by the following coordination graph over the K sensors:

Definition 4 (Coordination Graph) *Define the coordination graph to be an undirected graph $\mathcal{G}\triangleq(\mathcal{V},\mathcal{E})$ that comprises*

- *a set $\mathcal{V}$ of vertices denoting the K mobile sensors, and*
- *a set $\mathcal{E}$ of edges denoting coordination dependencies between sensors such that there exists an edge $\{k,k'\}$ incident with sensors $k\in\mathcal{V}$ and $k'\in\mathcal{V}\setminus\{k\}$ iff*

$$\max_{s\in Y_{W_k},s'\in Y_{W_{k'}}}\left|\Sigma_{sU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{Us'}\right| > \varepsilon \tag{19}$$

for a predefined constant $\varepsilon>0$ where W_k denotes the set of possible walks of length L of mobile sensor k from origin s_k in the road network G and $Y_{W_k}\triangleq\bigcup_{w_k\in W_k}Y_{w_k}$.

Remark. The construction of $\mathcal{G}$ can be decentralized as follows: Since $\ddot{\Sigma}_{UU}$ is symmetric and positive definite, it can be decomposed by Cholesky factorization into $\ddot{\Sigma}_{UU}=\Psi\Psi^\top$ where Ψ is a lower triangular matrix and $\Psi^\top$ is the transpose of Ψ. Then, $\Sigma_{sU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{Us'}=(\Psi\backslash\Sigma_{Us})^\top\Psi\backslash\Sigma_{Us'}$ where $\Psi\backslash B$ denotes the column vector ϕ solving $\Psi\phi=B$. That is, $\Sigma_{sU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{Us'}$ (19) can be expressed as a dot product of two vectors $\Psi\backslash\Sigma_{Us}$ and $\Psi\backslash\Sigma_{Us'}$; this property is exploited to determine adjacency between sensors in a decentralized manner:

Definition 5 (Adjacency) *Let*

$$\Phi_k\triangleq\{\Psi\backslash\Sigma_{Us}\}_{s\in Y_{W_k}} \tag{20}$$

for $k=1,\ldots,K$. A sensor $k\in\mathcal{V}$ is adjacent to sensor $k'\in\mathcal{V}\setminus\{k\}$ in coordination graph $\mathcal{G}$ iff

$$\max_{\phi\in\Phi_k,\phi'\in\Phi_{k'}}\left|\phi^\top\phi'\right|>\varepsilon \ . \tag{21}$$

It follows from the above definition that if each sensor k constructs Φ_k and exchanges it with every other sensor, then it can determine its adjacency to all the other sensors and store this information in a column vector a_k of length K with its k'-th component being defined as follows:

$$[a_k]_{k'}=\begin{cases}1 & \text{if sensor } k \text{ is adjacent to sensor } k',\\ 0 & \text{otherwise.}\end{cases} \tag{22}$$

By exchanging its adjacency vector a_k with every other sensor, each sensor can construct a globally consistent adjacency matrix $A_{\mathcal{G}}\triangleq(a_1\ldots a_K)$ to represent coordination graph $\mathcal{G}$.

Next, by computing the connected components (say, $\mathcal{K}$ of them) of coordination graph $\mathcal{G}$, their resulting vertex sets partition the set $\mathcal{V}$ of K sensors into $\mathcal{K}$ disjoint subsets $\mathcal{V}_1,\ldots,\mathcal{V}_{\mathcal{K}}$ such that the sensors within each subset have to coordinate their walks. Each sensor can determine its residing connected component in a decentralized way by performing a depth-first search in $\mathcal{G}$ starting from it as root.

Finally, construct a block-diagonal matrix $\widehat{\Sigma}_{Y_wY_w}$ to comprise diagonal blocks of the form $\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}$ for $n=1,\ldots,\mathcal{K}$ where $w_{\mathcal{V}_n}\triangleq(w_k)_{k\in\mathcal{V}_n}$ and $Y_{w_{\mathcal{V}_n}}\triangleq\bigcup_{k\in\mathcal{V}_n}Y_{w_k}$. The active sensing problem (17) is then approximated by

$$\begin{aligned}&\max_{w}\ \frac{1}{2}\log(2\pi e)^{|Y_w|}\left|\widehat{\Sigma}_{Y_wY_w}\right|\\ &\equiv\max_{(w_{\mathcal{V}_1},\ldots,w_{\mathcal{V}_{\mathcal{K}}})}\sum_{n=1}^{\mathcal{K}}\log(2\pi e)^{\left|Y_{w_{\mathcal{V}_n}}\right|}\left|\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}\right|\\ &=\sum_{n=1}^{\mathcal{K}}\max_{w_{\mathcal{V}_n}}\ \log(2\pi e)^{\left|Y_{w_{\mathcal{V}_n}}\right|}\left|\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}\right|,\end{aligned} \tag{23}$$

which can be solved in a partially decentralized manner by each disjoint subset $\mathcal{V}_n$ of mobile sensors:

$$\widehat{w}_{\mathcal{V}_n}=\arg\max_{w_{\mathcal{V}_n}}\ \log(2\pi e)^{\left|Y_{w_{\mathcal{V}_n}}\right|}\left|\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}\right| . \tag{24}$$

Our active sensing algorithm becomes fully decentralized if ε is set to be sufficiently large: more sensors become isolated in $\mathcal{G}$, consequently decreasing the size $\kappa\triangleq\max_{n}|\mathcal{V}_n|$

of its largest connected component to 1. As shown in Section 5.1, decreasing κ improves its time efficiency. On the other hand, it tends to a centralized behavior (17) by setting $\varepsilon \rightarrow 0^+$: $\mathcal{G}$ becomes near-complete, thus resulting in $\kappa \rightarrow K$.

Let
$$\xi \triangleq \max_{n, w_{\mathcal{V}_n}, i, i'} \left| \left[\left(\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}} Y_{w_{\mathcal{V}_n}}} \right)^{-1} \right]_{ii'} \right| \tag{25}$$
and $\epsilon \triangleq 0.5 \log 1 \Big/ \Big(1 - \big(K^{1.5} L^{2.5} \kappa \xi \varepsilon\big)^2\Big)$. In the result below, we prove that the joint walk $\widehat{w} \triangleq (\widehat{w}_{\mathcal{V}_1}, \ldots, \widehat{w}_{\mathcal{V}_K})$ is guaranteed to achieve an entropy $\overline{\mathbb{H}}[Z_{Y_{\widehat{w}}}]$ (i.e., by plugging $\widehat{w}$ into (18)) that is not more than ϵ from the maximum entropy $\overline{\mathbb{H}}[Z_{Y_{w^*}}]$ achieved by joint walk w^* (17):

Theorem 2 (Performance Guarantee) *If* $K^{1.5} L^{2.5} \kappa \xi \varepsilon < 1$, *then* $\overline{\mathbb{H}}[Z_{Y_{w^*}}] - \overline{\mathbb{H}}[Z_{Y_{\widehat{w}}}] \leq \epsilon$.

The proof of Theorem 2 is given in Appendix B. The implication of Theorem 2 is that our partially decentralized active sensing algorithm can perform comparatively well (i.e., small ϵ) under the following favorable environmental conditions: (a) the network of K sensors is not large, (b) length L of each sensor's walk to be optimized is not long, (c) the largest subset of κ sensors being formed to coordinate their walks (i.e., largest connected component in $\mathcal{G}$) is reasonably small, and (d) the minimum required correlation ε between walks of adjacent sensors is kept low.

Algorithm 1 below outlines the key operations of our D^{2}FAS algorithm to be run on each mobile sensor k, as detailed previously in Sections 3 and 4:

Algorithm 1: D^{2}FAS$(U, K, L, k, D_k, z_{D_k}, s_k)$

```
while true do
    /* Data fusion (Section 3)                                  */
    Construct local summary by (6) & (7)
    Exchange local summary with every sensor $i \neq k$
    Construct global summary by (8) & (9)
    Predict measurements at unobserved road segments by (10) & (11)
    /* Active Sensing (Section 4)                               */
    Construct $\Phi_k$ by (20)
    Exchange $\Phi_k$ with every sensor $i \neq k$
    Compute adjacency vector $a_k$ by (21) & (22)
    Exchange adjacency vector with every sensor $i \neq k$
    Construct adjacency matrix of coordination graph
    Find vertex set $\mathcal{V}_n$ of its residing connected component
    Compute maximum-entropy joint walk $\widehat{w}_{\mathcal{V}_n}$ by (24)
    Execute walk $\widehat{w}_k$ and observe its road segments $Y_{\widehat{w}_k}$
    Update local information $D_k$, $z_{D_k}$, and $s_k$ by (16)
```

5 Time and Communication Overheads

In this section, the time and communication overheads of our D^{2}FAS algorithm are analyzed and compared to that of centralized active sensing (17) coupled with the data fusion methods: Full GP (FGP) and SoD (Section 2).

5.1 Time Complexity

The data fusion component of D^{2}FAS involves computing the local and global summaries and the predictive Gaussian distribution. To construct the local summary using (6) and (7), each sensor has to evaluate $\Sigma_{D_k D_k|U}$ in $\mathcal{O}\big(|U|^3 + |U|(|D|/K)^2\big)$ time and invert it in $\mathcal{O}\big((|D|/K)^3\big)$ time, after which the local summary is obtained in $\mathcal{O}\big(|U|^2|D|/K + |U|(|D|/K)^2\big)$ time. The global summary is computed in $\mathcal{O}\big(|U|^2 K\big)$ by (8) and (9). Finally, the predictive Gaussian distribution is derived in $\mathcal{O}\big(|U|^3 + |U||Y|^2\big)$ time using (10) and (11). Supposing $|Y| \leq |U|$ for simplicity, the time complexity of data fusion is then $\mathcal{O}\big((|D|/K)^3 + |U|^3 + |U|^2 K\big)$.

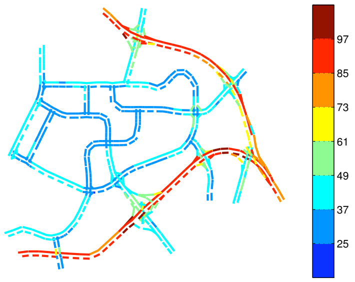
Let the maximum out-degree of G be denoted by δ. Then, each sensor has to consider $\Delta \triangleq \delta^L$ possible walks of length L. The active sensing component of D^{2}FAS involves computing Φ_k in $\mathcal{O}\big(\Delta L |U|^2\big)$ time, a_k in $\mathcal{O}\big(\Delta^2 L^2 |U| K\big)$ time, its residing connected component in $\mathcal{O}\big(\kappa^2\big)$ time, and the maximum-entropy joint walk by (11) and (24) with the following incurred time: The largest connected component of κ sensors in $\mathcal{G}$ has to consider Δ^κ possible joint walks. Note that $\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}} Y_{w_{\mathcal{V}_n}}} = \text{diag}\Big((\Sigma_{Y_{w_k} Y_{w_k}|U})_{k \in \mathcal{V}_n}\Big) + \Sigma_{Y_{w_{\mathcal{V}_n}} U} \ddot{\Sigma}_{UU}^{-1} \Sigma_{U Y_{w_{\mathcal{V}_n}}}$ where $\text{diag}(B)$ constructs a diagonal matrix by placing vector B on its diagonal. By exploiting Φ_k, the diagonal and latter matrix terms for all possible joint walks can be computed in $\mathcal{O}\big(\kappa\Delta(L|U|^2 + L^2|U|)\big)$ and $\mathcal{O}\big(\kappa^2 \Delta^2 L^2 |U|\big)$ time, respectively. For each joint walk $w_{\mathcal{V}_n}$, evaluating the determinant of $\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}} Y_{w_{\mathcal{V}_n}}}$ incurs $\mathcal{O}\big((\kappa L)^3\big)$ time. Therefore, the time complexity of active sensing is $\mathcal{O}\big(\kappa \Delta L |U|^2 + \Delta^2 L^2 |U| (K + \kappa^2) + \Delta^\kappa (\kappa L)^3\big)$.

Hence, the time complexity of our D^{2}FAS algorithm is $\mathcal{O}((|D|/K)^3 + |U|^2(|U| + K + \kappa\Delta L) + \Delta^2 L^2 |U| (K + \kappa^2) + \Delta^\kappa (\kappa L)^3)$. In contrast, the time incurred by centralized active sensing coupled with FGP and SoD are, respectively, $\mathcal{O}\big(|D|^3 + \Delta^K K L (|D|^2 + (KL)^2)\big)$ and $\mathcal{O}\big(|U|^3 |D| + \Delta^K K L (|U|^2 + (KL)^2)\big)$. It can be observed that D^{2}FAS can scale better with large $|D|$ (i.e., number of observations) and K (i.e., number of sensors). The scalability of D^{2}FAS vs. FGP and SoD will be further evaluated empirically in Section 6.

5.2 Communication Complexity

Let the communication overhead be defined as the size of each broadcast message. Recall from the data fusion component of D^{2}FAS in Algorithm 1 that, in each iteration, each sensor broadcasts a $\mathcal{O}\big(|U|^2\big)$-sized summary encapsulating its local observations, which is robust against communication failure. In contrast, FGP and SoD require each sensor to broadcast, in each iteration, a $\mathcal{O}(|D|/K)$-sized message comprising exactly its local observations to handle communication failure. If the number of local observations grows to be larger in size than a local summary of predefined size, then the data fusion component of D^{2}FAS is more scalable than FGP and SoD in terms of communication overhead. For the partially decentralized active sensing component of D^{2}FAS, each sensor broadcasts $\mathcal{O}(\Delta L |U|)$-sized Φ_k and $\mathcal{O}(K)$-sized a_k messages.

6 Experiments and Discussion



This section evaluates the predictive performance, time efficiency, and scalability of our D^{2}FAS algorithm on a real-world traffic phenomenon (i.e., speeds (km/h) of road segments) over an urban road network (top figure) in Tampines area, Singapore during evening peak hours on April 20, 2011. It comprises 775 road segments including highways, arterials, slip roads, etc. The mean speed is 48.8 km/h and the standard deviation is 20.5 km/h.

The performance of D^{2}FAS is compared to that of centralized active sensing (17) coupled with the state-of-the-art data fusion methods: full GP (FGP) and SoD (Section 2). A network of K mobile sensors is tasked to explore the road network to gather a total of up to 960 observations. To reduce computational time, each sensor repeatedly computes and executes maximum-entropy walks of length $L = 2$ (instead of computing a very long walk), unless otherwise stated. For D^{2}FAS and SoD, $|U|$ is set to 64 . For the active sensing component of D^{2}FAS, ε is set to 0.1, unless otherwise stated. The experiments are run on a Linux PC with Intel® Core$^{\text{TM}}$2 Quad CPU Q9550 at 2.83 GHz.

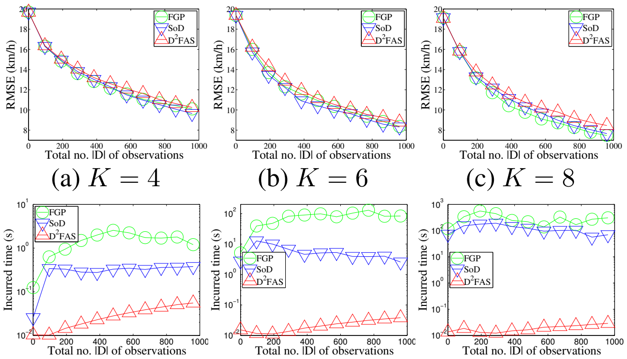
6.1 Performance Metrics

The first metric evaluates the predictive performance of a tested algorithm: It measures the *root mean squared error* (RMSE) $\sqrt{|V|^{-1}\sum_{s\in V}(z_s - \widehat{\mu}_s)^2}$ over the entire domain V of the road network that is incurred by the predictive mean $\widehat{\mu}_s$ of the tested algorithm, specifically, using (1) of FGP, (4) of SoD, or (10) of D^{2}FAS. The second metric evaluates the time efficiency and scalability of a tested algorithm by measuring its incurred time; for D^{2}FAS, the maximum of the time incurred by all subsets $\mathcal{V}_1, \ldots, \mathcal{V}_{\mathcal{K}}$ of sensors is recorded.

6.2 Results and Analysis

Predictive performance and time efficiency. Fig. 1 shows results of the performance of the tested algorithms averaged over 40 randomly generated starting sensor locations with varying number $K = 4, 6, 8$ of sensors. It can be observed that D^{2}FAS is significantly more time-efficient and scales better with increasing number $|D|$ of observations (Figs. 1d to 1f) while achieving predictive performance close to that of centralized active sensing coupled with FGP and SoD (Figs. 1a to 1c). Specifically, D^{2}FAS is about $1, 2, 4$ orders of magnitude faster than centralized active sensing coupled with FGP and SoD for $K = 4, 6, 8$ sensors, respectively.

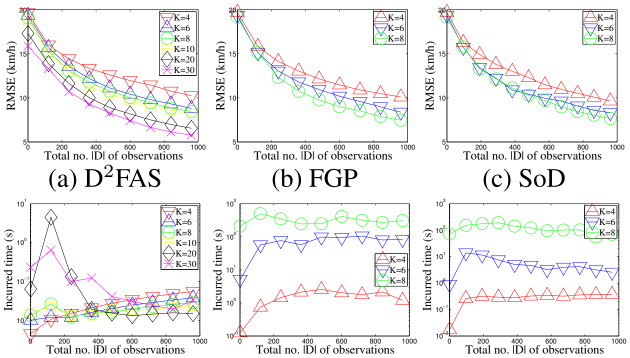
Scalability of D^{2}FAS. Using the same results as that in Fig. 1, Fig. 2 plots them differently to reveal the scalability of the tested algorithms with increasing number K of sensors. Additionally, we provide results of the performance of D^{2}FAS for $K = 10, 20, 30$ sensors; such results are not available for centralized active sensing coupled with FGP and SoD due to extremely long incurred time. It can be observed from Figs. 2a to 2c that the predictive performance of all tested algorithms improve with a larger number of sensors because each sensor needs to execute fewer walks and its performance is therefore less adversely affected by its myopic selection (i.e., $L = 2$) of maximum-entropy walks. As a result, more informative unobserved road segments are explored.



(a) $K = 4$ (b) $K = 6$ (c) $K = 8$

(d) $K = 4$ (e) $K = 6$ (f) $K = 8$

Figure 1: Graphs of (a-c) predictive performance and (d-f) time efficiency vs. total no. $|D|$ of observations gathered by varying number K of mobile sensors.

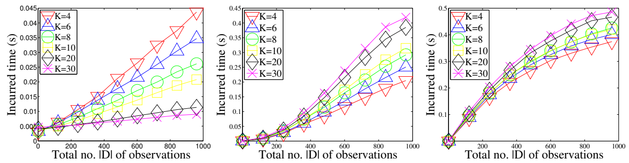


(a) D^{2}FAS (b) FGP (c) SoD

(d) D^{2}FAS (e) FGP (f) SoD

Figure 2: Graphs of (a-c) predictive performance and (d-f) time efficiency vs. total no. $|D|$ of observations gathered by varying number K of mobile sensors.

As shown in Fig. 2d, when the randomly placed sensors gather their initial observations (i.e., $|D| < 400$), the time incurred by D^{2}FAS is higher for greater K due to larger subsets of sensors being formed to coordinate their walks (i.e., larger κ). As more observations are gathered (i.e., $|D| \geq 400$), its partially decentralized active sensing component directs the sensors to explore further apart from each other in order to maximize the entropy of their walks. This consequently decreases κ, leading to a reduction in incurred time. Furthermore, as K increases from 4 to 20, the incurred time decreases due to its decentralized data fusion component that can distribute the computational load



(a) D^{2}FAS (b) FGP (c) SoD

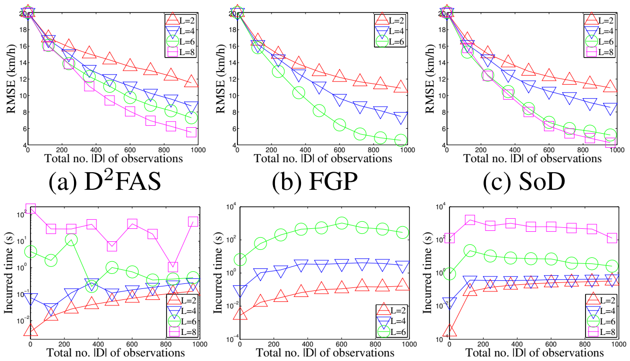
Figure 3: Graphs of time efficiency vs. total no. $|D|$ of observations gathered by varying number K of sensors.

among a greater number of sensors. When the road network becomes more crowded from $K = 20$ to $K = 30$ sensors, the incurred time increases slightly due to slightly larger κ. In contrast, Figs. 2e and 2f show that the time taken by FGP and SoD increases significantly primarily due to their centralized active sensing incurring exponential time in K. Hence, the scalability of our D^{2}FAS algorithm in the number of sensors allows the deployment of a larger-scale mobile sensor network (i.e., $K \geq 10$) to achieve more accurate traffic modeling and prediction (Figs. 2a to 2c).

Scalability of data fusion. Fig. 3 shows results of the scalability of the tested data fusion methods with increasing number K of sensors. In order to produce meaningful results for fair comparison, the same active sensing component has to be coupled with the data fusion methods and its incurred time kept to a minimum. As such, we impose the use of a fully decentralized active sensing component to be performed by each mobile sensor k: $w_k^* = \arg\max_{w_k} \mathbb{H}[Z_{Y_{w_k}}|Z_D]$. For D^{2}FAS, this corresponds exactly to (24) by setting a large enough ε (in our experiments, $\varepsilon = 2$) to yield $\kappa = 1$; consequently, computational and communicational operations pertaining to the coordination graph can be omitted.

It can be seen from Fig. 3a that the time incurred by the decentralized data fusion component of D^{2}FAS decreases with increasing K, as explained previously. In contrast, the time incurred by FGP and SoD increases (Fig. 3b and 3c): As discussed above, a larger number of sensors results in a greater quantity of more informative unique observations to be gathered (i.e., fewer repeated observations), which increases the time needed for data fusion. When $K \geq 10$, D^{2}FAS is at least 1 order of magnitude faster than FGP and SoD. It can also be observed that D^{2}FAS scales better with increasing number of observations. So, the real-time performance and scalability of D^{2}FAS's decentralized data fusion enable it to be used for persistent large-scale traffic modeling and prediction where a large number of observations and sensors (including static and passive ones) are expected to be available.

Varying length L of walk. Fig. 4 shows results of the performance of the tested algorithms with varying length $L = 2, 4, 6, 8$ of maximum-entropy joint walks; we choose to experiment with just 2 sensors since Figs. 2 and 3 reveal that a smaller number of sensors produce poorer predictive performance and higher incurred time with large number of observations for D^{2}FAS. It can be observed that the predictive performance of all tested algorithms improve with increasing walk length L because the selection of maximum-entropy joint walks is less myopic. The time incurred by D^{2}FAS increases due to larger κ but grows more slowly and is lower than that incurred by centralized active sensing coupled with FGP and SoD. Specifically, when $L = 8$, D^{2}FAS is at least 1 order of magnitude faster (i.e., average of 60 s) than centralized active sensing coupled with SoD (i.e., average of > 732 s) and FGP (i.e., not available due to excessive incurred time). Also, notice from Figs. 2a and 2d that if a large number of sensors (i.e., $K = 30$) is available, D^{2}FAS can select shorter walks of $L = 2$ to be significantly more time-efficient (i.e., average of > 3 orders of magnitude faster) while achieving predictive performance comparable to that of SoD with $L = 8$ and FGP with $L = 6$.



(a) D^{2}FAS (b) FGP (c) SoD

(d) D^{2}FAS (e) FGP (f) SoD

Figure 4: Graphs of (a-c) predictive performance and (d-f) time efficiency vs. total no. $|D|$ of observations gathered by 2 mobile sensors with varying length L of maximum-entropy joint walks.

7 Conclusion

This paper describes a decentralized data fusion and active sensing algorithm for modeling and predicting spatiotemporal traffic phenomena with mobile sensors. Analytical and empirical results have shown that our D^{2}FAS algorithm is significantly more time-efficient and scales better with increasing number of observations and sensors while achieving predictive performance close to that of state-of-the-art centralized active sensing coupled with FGP and SoD. Hence, D^{2}FAS is practical for deployment in a large-scale mobile sensor network to achieve persistent and accurate traffic modeling and prediction. For our future work, we will assume that each sensor can only communicate locally with its neighbors (instead of assuming all-to-all communication) and develop a *distributed* data fusion approach to efficient and scalable approximate GP prediction based on D^{2}FAS and consensus filters (Olfati-Saber, 2005).

Acknowledgments. This work was supported by Singapore-MIT Alliance Research and Technology (SMART) Subaward Agreement 14 R-252-000-466-592.

References

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer, NY.

Chen, H., Rakha, H. A., and Sadek, S. (2011). Real-time freeway traffic state prediction: A particle filter approach. In *Proc. IEEE ITSC*, pages 626–631.

Chu, C.-T., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G. R., Ng, A. Y., and Olukotun, K. (2007). Map-Reduce for machine learning on multicore. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 281–288, Cambridge, MA. MIT Press.

Chung, T. H., Gupta, V., Burdick, J. W., and Murray, R. M. (2004). On a decentralized active sensing strategy using mobile sensor platforms in a network. In *Proc. CDC*, pages 1914–1919.

Coates, M. (2004). Distributed particle filters for sensor networks. In *Proc. IPSN*, pages 99–107.

Cortes, J. (2009). Distributed kriged Kalman filter for spatial estimation. *IEEE Trans. Automat. Contr.*, **54**(12), 2816–2827.

Dolan, J. M., Podnar, G., Stancliff, S., Low, K. H., Elfes, A., Higinbotham, J., Hosler, J. C., Moisan, T. A., and Moisan, J. (2009). Cooperative aquatic sensing using the telesupervised adaptive ocean sensor fleet. In *Proc. SPIE Conference on Remote Sensing of the Ocean, Sea Ice, and Large Water Regions*, volume 7473.

Golub, G. H. and Van Loan, C.-F. (1996). *Matrix Computations*. Johns Hopkins Univ. Press, third edition.

Guestrin, C., Bodik, P., Thibaus, R., Paskin, M., and Madden, S. (2004). Distributed regression: an efficient framework for modeling sensor network data. In *Proc. IPSN*, pages 1–10.

Ipsen, I. C. F. and Lee, D. J. (2003). Determinant approximations. Technical Report CRSC-TR03-30, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC.

Kamarianakis, Y. and Prastacos, P. (2003). Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transport. Res. Rec.*, **1857**, 74–84.

Krause, A., Horvitz, E., Kansal, A., and Zhao, F. (2008a). Toward community sensing. In *Proc. IPSN*, pages 481–492.

Krause, A., Singh, A., and Guestrin, C. (2008b). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, **9**, 235–284.

Lawrence, N., Seeger, M., and Herbrich, R. (2003). Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616, Cambridge, MA. MIT Press.

Low, K. H., Gordon, G. J., Dolan, J. M., and Khosla, P. (2007). Adaptive sampling for multi-robot wide-area exploration. In *Proc. IEEE ICRA*, pages 755–760.

Low, K. H., Dolan, J. M., and Khosla, P. (2008). Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS*, pages 23–30.

Low, K. H., Dolan, J. M., and Khosla, P. (2009a). Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*, pages 233–240.

Low, K. H., Podnar, G., Stancliff, S., Dolan, J. M., and Elfes, A. (2009b). Robot boats as a mobile aquatic sensor network. In *Proc. IPSN-09 Workshop on Sensor Networks for Earth and Space Science Applications*.

Low, K. H., Dolan, J. M., and Khosla, P. (2011). Active Markov information-theoretic path planning for robotic environmental sensing. In *Proc. AAMAS*, pages 753–760.

Low, K. H., Chen, J., Dolan, J. M., Chien, S., and Thompson, D. R. (2012). Decentralized active robotic exploration and mapping for probabilistic field classification in environmental sensing. In *Proc. AAMAS*, pages 105–112.

Min, W. and Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transport. Res. C-Emer.*, **19**(4), 606–616.

Neumann, M., Kersting, K., Xu, Z., and Schulz, D. (2009). Stacked Gaussian process learning. In *Proc. ICDM*, pages 387–396.

Oh, S., Xu, Y., and Choi, J. (2010). Explorative navigation of mobile sensor networks using sparse Gaussian processes. In *Proc. CDC*, pages 3851–3856.

Olfati-Saber, R. (2005). Distributed Kalman filter with embedded consensus filters. In *Proc. CDC*, pages 8179–8184.

Paskin, M. A. and Guestrin, C. (2004). Robust probabilistic inference in distributed systems. In *Proc. UAI*, pages 436–445.

Podnar, G., Dolan, J. M., Low, K. H., and Elfes, A. (2010). Telesupervised remote surface water quality sensing. In *Proc. IEEE Aerospace Conference*.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *JMLR*, **6**, 1939–1959.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Rosencrantz, M., Gordon, G., and Thrun, S. (2003). Decentralized sensor fusion with distributed particle filters. In *Proc. UAI*, pages 493–500.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, first edition.

Schrank, D., Lomax, T., and Eisele, B. (2011). *TTI's 2011 Urban Mobility Report*. Texas Transportation Institute, Texas A&M University.

Seeger, M. and Williams, C. (2003). Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proc. AISTATS*.

Srinivasan, K. K. and Jovanis, P. P. (1996). Determination of number of probe vehicle required for reliable travel time measurement in urban network. *Transport. Res. Rec.*, **1537**, 15–22.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory*. Academic Press.

Stranders, R., Farinelli, A., Rogers, A., and Jennings, N. R. (2009). Decentralised coordination of mobile sensors using the max-sum algorithm. In *Proc. IJCAI*, pages 299–304.

Sukkarieh, S., Nettleton, E., Kim, J., Ridley, M., Goktogan, A., and Durrant-Whyte, H. (2003). The ANSER project: Data fusion across multiple uninhabited air vehicles. *IJRR*, **22**(7-8), 505–539.

Turner, S. M., Eisele, W. L., Benz, R. J., and Holdener, D. J. (1998). Travel time data collection handbook. Technical Report FHWA-PL-98-035, Federal Highway Administration, Office of Highway Information Management, Washington, DC.

Wang, Y. and Papageorgiou, M. (2005). Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transport. Res. B-Meth.*, **39**(2), 141–167.

Work, D. B., Blandin, S., Tossavainen, O., Piccoli, B., and Bayen, A. (2010). A traffic model for velocity data assimilation. *AMRX*, **2010**(1), 1–35.

A Proof of Theorem 1B

We have to first simplify the $\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}$ term in the expressions of $\mu^{\text{PITC}}_{Y|D}$ (12) and $\Sigma^{\text{PITC}}_{YY|D}$ (13).

$$\begin{aligned}
&\left(\Gamma_{DD}+\Lambda\right)^{-1}\\
&=\left(\Sigma_{DU}\Sigma^{-1}_{UU}\Sigma_{UD}+\Lambda\right)^{-1}\\
&=\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\left(\Sigma_{UU}+\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\right)^{-1}\Sigma_{UD}\Lambda^{-1}\\
&=\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\ .
\end{aligned}\tag{26}$$

The second equality follows from matrix inversion lemma. The last equality is due to

$$\begin{aligned}
&\Sigma_{UU}+\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\\
&=\Sigma_{UU}+\sum_{k=1}^{K}\Sigma_{UD_k}\Sigma^{-1}_{D_kD_k|U}\Sigma_{D_kU}\\
&=\Sigma_{UU}+\sum_{k=1}^{K}\dot{\Sigma}^{k}_{UU}=\ddot{\Sigma}_{UU}\ .
\end{aligned}\tag{27}$$

Using (14) and (26),

$$\begin{aligned}
&\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\\
&=\Sigma_{YU}\Sigma^{-1}_{UU}\Sigma_{UD}\left(\Lambda^{-1}-\Lambda^{-1}\Sigma_{DU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\right)\\
&=\Sigma_{YU}\Sigma^{-1}_{UU}\left(\ddot{\Sigma}_{UU}-\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\right)\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\\
&=\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}
\end{aligned}\tag{28}$$

The third equality is due to (27).

From (12),

$$\begin{aligned}
\mu^{\text{PITC}}_{Y|D}&=\mu_Y+\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\left(z_D-\mu_D\right)\\
&=\mu_Y+\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\left(z_D-\mu_D\right)\\
&=\mu_Y+\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\ddot{z}_U\\
&=\overline{\mu}_Y\ .
\end{aligned}$$

The second equality is due to (28). The third equality follows from $\Sigma_{UD}\Lambda^{-1}\left(z_D-\mu_D\right)=\sum_{k=1}^{K}\Sigma_{UD_k}\Sigma^{-1}_{D_kD_k|U}\left(z_{D_k}-\mu_{D_k}\right)=\sum_{k=1}^{K}\dot{z}^k_U=\ddot{z}_U$.

From (13),

$$\begin{aligned}
&\Sigma^{\text{PITC}}_{YY|D}\\
&=\Sigma_{YY}-\Gamma_{YD}\left(\Gamma_{DD}+\Lambda\right)^{-1}\Gamma_{DY}\\
&=\Sigma_{YY}-\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\Sigma^{-1}_{UU}\Sigma_{UY}\\
&=\Sigma_{YY}-\Big(\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}\Sigma^{-1}_{UU}\Sigma_{UY}\\
&\quad-\Sigma_{YU}\Sigma^{-1}_{UU}\Sigma_{UY}\Big)-\Sigma_{YU}\Sigma^{-1}_{UU}\Sigma_{UY}\\
&=\Sigma_{YY}-\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\left(\Sigma_{UD}\Lambda^{-1}\Sigma_{DU}-\ddot{\Sigma}_{UU}\right)\Sigma^{-1}_{UU}\Sigma_{UY}\\
&\quad-\Sigma_{YU}\Sigma^{-1}_{UU}\Sigma_{UY}\\
&=\Sigma_{YY}-\left(\Sigma_{YU}\Sigma^{-1}_{UU}\Sigma_{UY}-\Sigma_{YU}\ddot{\Sigma}^{-1}_{UU}\Sigma_{UY}\right)\\
&=\Sigma_{YY}-\Sigma_{YU}\left(\Sigma^{-1}_{UU}-\ddot{\Sigma}^{-1}_{UU}\right)\Sigma_{UY}\\
&=\overline{\Sigma}_{YY}\ .
\end{aligned}$$

The second equality follows from (14) and (28). The fifth equality is due to (27).

B Proof of Theorem 2

Let $\widetilde{\Sigma}_{Y_wY_w}\triangleq\overline{\Sigma}_{Y_wY_w}-\widehat{\Sigma}_{Y_wY_w}$ and ρ_w be the spectral radius of $\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}$. We have to first bound ρ_w from above.

For any joint walk w, $\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}$ comprises diagonal blocks of size $\left|Y_{w_{\mathcal{V}_n}}\right|\times\left|Y_{w_{\mathcal{V}_n}}\right|$ with components of value 0 for $n=1,\ldots,\mathcal{K}$ and off-diagonal blocks of the form $\left(\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}\right)^{-1}\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_{n'}}}}$ for $n,n'=1,\ldots,\mathcal{K}$ and $n\neq n'$. We know that any pair of sensors $k\in\mathcal{V}_n$ and $k'\in\mathcal{V}_{n'}$ reside in different connected components of coordination graph $\mathcal{G}$ and are therefore not adjacent. So, by Definition 5,

$$\max_{i,i'}\left|\left[\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_{n'}}}}\right]_{ii'}\right|\leq\varepsilon\tag{29}$$

for $n,n'=1,\ldots,\mathcal{K}$ and $n\neq n'$. Using (25) and (29), each component in any off-diagonal block of $\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}$ can be bounded as follows:

$$\max_{i,i'}\left|\left[\left(\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_n}}}\right)^{-1}\overline{\Sigma}_{Y_{w_{\mathcal{V}_n}}Y_{w_{\mathcal{V}_{n'}}}}\right]_{ii'}\right|\leq\left|Y_{w_{\mathcal{V}_n}}\right|\xi\varepsilon\tag{30}$$

for $n,n'=1,\ldots,\mathcal{K}$ and $n\neq n'$. It follows from (30) that

$$\max_{i,i'}\left|\left[\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}\right]_{ii'}\right|\leq\max_{n}\left|Y_{w_{\mathcal{V}_n}}\right|\xi\varepsilon\leq L\kappa\xi\varepsilon\ .\tag{31}$$

The last inequality is due to $\max\limits_{n}\left|Y_{w_{\mathcal{V}_n}}\right|\leq L\max\limits_{n}|\mathcal{V}_n|\leq L\kappa$. Then,

$$\begin{aligned}
\rho_w&\leq\left\|\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}\right\|_2\\
&\leq|Y_w|\max_{i,i'}\left|\left[\left(\widehat{\Sigma}_{Y_wY_w}\right)^{-1}\widetilde{\Sigma}_{Y_wY_w}\right]_{ii'}\right|\\
&\leq KL^2\kappa\xi\varepsilon\ .
\end{aligned}\tag{32}$$

The first two inequalities follow from standard properties of matrix norm (Golub and Van Loan, 1996; Stewart and Sun, 1990). The last inequality is due to (31).

The rest of this proof utilizes the following result of Ipsen and Lee (2003) that is revised to reflect our notations:

Theorem 3 *If* $|Y_w|\rho^2_w<1$, *then* $\log\left|\overline{\Sigma}_{Y_wY_w}\right|\leq\log\left|\widehat{\Sigma}_{Y_wY_w}\right|\leq\log\left|\overline{\Sigma}_{Y_wY_w}\right|-\log\left(1-|Y_w|\rho^2_w\right)$ *for any joint walk* w.

Using Theorem 3 followed by (32),

$$\begin{aligned}
\log\left|\widehat{\Sigma}_{Y_wY_w}\right|-\log\left|\overline{\Sigma}_{Y_wY_w}\right|&\leq\log\frac{1}{1-|Y_w|\rho^2_w}\\
&\leq\log\frac{1}{1-\left(K^{1.5}L^{2.5}\kappa\xi\varepsilon\right)^2}
\end{aligned}\tag{33}$$

for any joint walk w.

$$
\begin{aligned}
&\overline{\mathbb{H}}[Z_{Y_{w^*}}] - \overline{\mathbb{H}}[Z_{Y_{\widehat{w}}}] \\
&= \frac{1}{2}\left((|Y_{w^*}| - |Y_{\widehat{w}}|)\log(2\pi e) + \log\left|\overline{\Sigma}_{Y_{w^*}Y_{w^*}}\right| - \log\left|\overline{\Sigma}_{Y_{\widehat{w}}Y_{\widehat{w}}}\right|\right) \\
&\leq \frac{1}{2}\left((|Y_{w^*}| - |Y_{\widehat{w}}|)\log(2\pi e) + \log\left|\widehat{\Sigma}_{Y_{w^*}Y_{w^*}}\right| - \log\left|\overline{\Sigma}_{Y_{\widehat{w}}Y_{\widehat{w}}}\right|\right) \\
&\leq \frac{1}{2}\left((|Y_{\widehat{w}}| - |Y_{\widehat{w}}|)\log(2\pi e) + \log\left|\widehat{\Sigma}_{Y_{\widehat{w}}Y_{\widehat{w}}}\right| - \log\left|\overline{\Sigma}_{Y_{\widehat{w}}Y_{\widehat{w}}}\right|\right) \\
&\leq \frac{1}{2}\log\frac{1}{1-(K^{1.5}L^{2.5}\kappa\xi\varepsilon)^2}\ .
\end{aligned}
$$

The first equality is due to (18). The first, second, and last inequalities follow from Theorem 3, (23), and (33), respectively.