Gaussian Process Decentralized Data Fusion Meets Transfer Learning in Large-Scale Distributed Cooperative Perception

Ruofei Ouyang and **Kian Hsiang Low**
Department of Computer Science
National University of Singapore, Republic of Singapore
{ouyang, lowkh}@comp.nus.edu.sg

Abstract

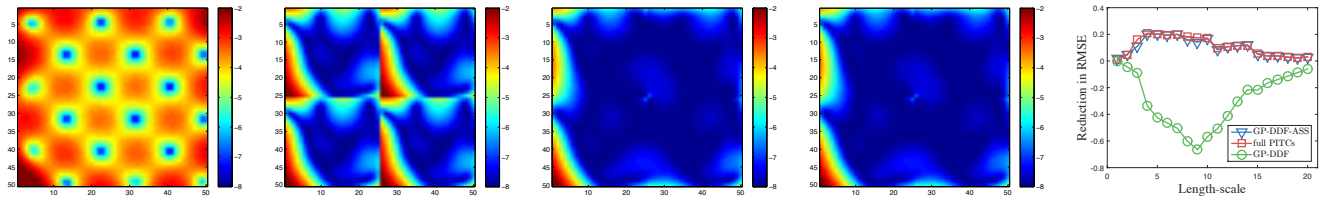
This paper presents novel Gaussian process decentralized data fusion algorithms exploiting the notion of agent-centric support sets for distributed cooperative perception of large-scale environmental phenomena. To overcome the limitations of scale in existing works, our proposed algorithms allow every mobile sensing agent to choose a different support set and dynamically switch to another during execution for encapsulating its own data into a local summary that, perhaps surprisingly, can still be assimilated with the other agents' local summaries (i.e., based on their current choices of support sets) into a globally consistent summary to be used for predicting the phenomenon. To achieve this, we propose a novel transfer learning mechanism for a team of agents capable of sharing and transferring information encapsulated in a summary based on a support set to that utilizing a different support set with some loss that can be theoretically bounded and analyzed. To alleviate the issue of information loss accumulating over multiple instances of transfer learning, we propose a new information sharing mechanism to be incorporated into our algorithms in order to achieve memory-efficient lazy transfer learning. Empirical evaluation on real-world datasets show that our algorithms outperform the state-of-the-art methods.

1 Introduction

Central to many environmental sensing and monitoring applications (e.g., traffic flow and mobility demand predictions over urban road networks (Chen et al. 2015), monitoring of ocean and freshwater phenomena (Dolan et al. 2009), adaptive sampling and active sensing/learning (Cao, Low, and Dolan 2013; Hoang et al. 2014; Low, Dolan, and Khosla 2008; 2009; 2011; Low et al. 2007; 2012; Ouyang et al. 2014; Zhang et al. 2016), Bayesian optimization (Daxberger and Low 2017; Hoang, Hoang, and Low 2018; Ling, Low, and Jaillet 2016), among others) is the need to scale up data fusion algorithms for big data because massive volumes of data/observations gathered by multiple static and/or mobile sensing agents have to be assimilated to form a globally consistent predictive belief of the environmental phenomenon of interest. A centralized approach to data fusion is ill-suited here because it suffers from poor scalability in the data size and a single point of failure. To this end, decentralized data fusion algorithms such as distributed Bayesian filtering (Olfati-Saber 2005) and distributed regression (Guestrin et al. 2004) have been developed to improve scalability and robustness to failure.

Recent works (Chen et al. 2012; 2015; Chen, Low, and Tan 2013; Cortes 2009) have progressed from the use of simple Markov parametric models assuming independent observations (e.g., in distributed Bayesian filtering) to that of a rich class of Bayesian nonparametric *Gaussian process* (GP) models characterizing continuous-valued, spatially correlated observations in order to represent the latent structure of the spatially varying, possibly noisy phenomenon with higher fidelity. Instead of communicating the local data of each sensing agent directly to every other agent which is not scalable, the *GP decentralized data fusion* (GP-DDF) algorithms of Chen et al. (2015) enable the agents to encapsulate their own data into constant-sized local summaries, exchange them, and finally assimilate them into a globally consistent summary to be exploited for predicting the phenomenon. Different from the above distributed regression algorithms, they do not need to exploit spatial locality assumptions for gaining efficiency and can thus be used for mobile sensing agents whose paths are not constrained by locality. They also do not suffer from the drawbacks of the GP distributed data fusion algorithm of Cortes (2009) relying on an iterative procedure of weighted least squares, which assumes bounded correlation and uncorrelated past observations that can severely compromise its predictive performance and converges very slowly in the case of a large number of agents. In contrast, the GP-DDF algorithms can be computed exactly and efficiently. More importantly, their predictive performance can be theoretically guaranteed to be equivalent to that of sophisticated centralized sparse approximations (Chen et al. 2013; Hoang, Hoang, and Low 2015; 2016; 2017; Low et al. 2015; Quiñonero-Candela and Rasmussen 2005; Snelson and Ghahramani 2007; Xu et al. 2014) of the GP model.

However, like their centralized counterparts, the GP-DDF algorithms rely on the notion of a *fixed support set* of input locations *common* to all agents for encapsulating their own data into local summaries, which raises three non-trivial issues limiting their scalability to small domains of spatial phenomena and hence small data sizes: (a) When the domain is expanded, the support set must be increased propor-



(a) GP-DDF (b) Local PITCs (c) Full PITCs (d) GP-DDF-ASS (e)

Figure 1: (a-d) Maps of log-predictive variance/uncertainty (i.e., $\log \overline{\sigma}_x^2$ (5) for all $x \in \mathcal{X}$) over a simulated spatial phenomenon with length-scale of 10 achieved by various decentralized data fusion algorithms given the same data and support set size for each agent, and (e) graphs of reduction in RMSE of GP-DDF, full PITCs, and GP-DDF-ASS over local PITCs vs. varying length-scales. Experimental setup, results, and analysis for this simulated experiment are detailed in Section 4.1.

tionally in size to cover and predict the phenomenon well at the expense of greater time, space, and communication overheads, which grows prohibitively costly; (b) supposing the support set is restricted in size to limit the overheads and thus only sparsely covers the large-scale phenomenon, huge information loss due to summarization (and consequently high predictive uncertainty, as shown in Fig. 1a) is expected, especially when the local data gathered by the possibly "close" agents are "far" (i.e., in the correlation sense) from the support set; and (c) if the current support set needs to be replaced by a new support set of different size and input locations (e.g., due to change in domain size or time, space, and communication requirements, using an improved active learning criterion to select a support set that better covers and predicts the phenomenon), then all previously gathered data (if not already discarded after summarization using old support set) have to be re-encapsulated into local summaries based on the new support set, which is not scalable.

To address these challenging issues faced by GP-DDF algorithms, this paper presents novel Gaussian process decentralized data fusion algorithms with *agent-centric support sets* (Section 3) for distributed cooperative perception of large-scale environmental phenomena. In contrast to existing GP-DDF algorithms, our proposed algorithms allow every sensing agent to choose a possibly different support set and dynamically switch to another during execution for encapsulating its own data into a local summary that, perhaps surprisingly, can still be assimilated with the other agents' local summaries (i.e., based on their current choices of support sets) into a globally consistent summary to be used for predicting the phenomenon. To achieve this, we propose a novel *transfer learning* mechanism for a team of mobile sensing agents capable of sharing and transferring information encapsulated in a summary based on a support set to that utilizing a different support set with some loss that can be theoretically bounded and analyzed, which is the *main contribution* of our work here. To alleviate the issue of information loss accumulating over multiple instances of transfer learning, we propose a new *information sharing* mechanism to be incorporated into our GP-DDF algorithms with agent-centric support sets in order to achieve *memory-efficient lazy transfer learning*. As a result, our algorithms can resolve the above-mentioned critical issues plaguing existing GP-DDF algorithms: (a) For any unobserved input location, an agent can choose a small, constant-sized (i.e., independent of domain size of phenomenon) but sufficiently dense support set surrounding it to predict its measurement accurately with much lower predictive uncertainty (see Fig. 1d) while preserving time, space, and communication efficiencies; (b) the agents can reduce the information loss due to summarization by choosing or dynamically switching to a support set "close" to their local data; and (c) without needing to retain previously gathered data, an agent can choose or dynamically switch to a new support set whose summary can be constructed using information transferred from the summary based on its current support set, thus preserving scalability to big data. We empirically evaluate the performance of our algorithms using real-world datasets featuring indoor lighting quality gathered by a team of 3 real Pioneer 3-DX mobile robots and sea surface temperature of the Indian ocean explored by 64 agents; the latter is millions in size (Section 4).

2 Background and Notations

Modeling Spatially Varying Environmental Phenomena with Gaussian Processes (GPs). A GP can model a spatially varying environmental phenomenon as follows: The phenomenon is defined to vary as a realization of a GP. Let $\mathcal{X}$ be a set representing the domain of the phenomenon such that each location $x \in \mathcal{X}$ is associated with a realized (random) measurement y_x (Y_x) if it is observed (unobserved). Let $\{Y_x\}_{x\in\mathcal{X}}$ denote a GP, that is, any finite subset of $\{Y_x\}_{x\in\mathcal{X}}$ follows a multivariate Gaussian distribution. Then, the GP is fully specified by its *prior* mean $\mu_x \triangleq \mathbb{E}[Y_x]$ and covariance $\sigma_{xx'} \triangleq \mathrm{cov}[Y_x, Y_{x'}]$ for all $x, x' \in \mathcal{X}$, the latter of which characterizes the spatial correlation structure of the phenomenon and can be defined, for example, by the squared exponential covariance function

$$\sigma_{xx'} \triangleq \sigma_s^2 \exp(-0.5\|\Lambda^{-1}(x - x')\|^2) + \sigma_n^2\delta_{xx'} \qquad (1)$$

where σ_s^2 (σ_n^2) is its signal (noise) variance hyperparameter controlling the intensity (noise) of the measurements, Λ is a diagonal matrix with length-scale hyperparameters ℓ_1 and ℓ_2 controlling, respectively, the degree of spatial correlation or "similarity" between measurements in the horizontal and vertical directions of the phenomenon, and $\delta_{xx'}$ is a Kronecker delta that is 1 if $x = x'$, and 0 otherwise.

Supposing a column vector $y_\mathcal{D} \triangleq (y_{x'})^\top_{x'\in\mathcal{D}}$ of realized measurements is observed for some set $\mathcal{D} \subset \mathcal{X}$ of loca-

tions, a GP model can exploit these observations/data to perform probabilistic regression by providing a Gaussian *posterior*/predictive distribution

$$\mathcal{N}(\mu_x + \Sigma_{x\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}(y_{\mathcal{D}} - \mu_{\mathcal{D}}), \sigma_{xx} - \Sigma_{x\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}x}) \quad (2)$$

of the measurement for any unobserved location $x \in \mathcal{X} \setminus \mathcal{D}$ where $\mu_{\mathcal{D}} \triangleq (\mu_{x'})_{x' \in \mathcal{D}}^{\top}$, $\Sigma_{x\mathcal{D}} \triangleq (\sigma_{xx'})_{x' \in \mathcal{D}}$, $\Sigma_{\mathcal{D}\mathcal{D}} \triangleq (\sigma_{x'x''})_{x', x'' \in \mathcal{D}}$, and $\Sigma_{\mathcal{D}x} \triangleq \Sigma_{x\mathcal{D}}^{\top}$. To predict the phenomenon, a naive approach to data fusion is to fully communicate all the data to every mobile sensing agent, each of which then predicts the phenomenon separately using the Gaussian predictive distribution in (2). Such an approach, however, scales poorly in the data size $|\mathcal{D}|$ due to the need to invert $\Sigma_{\mathcal{D}\mathcal{D}}$ which incurs $\mathcal{O}(|\mathcal{D}|^3)$ time.

GP Decentralized Data Fusion (GP-DDF). To improve the scalability of the GP model for practical use in data fusion, the work of Chen et al. (2015) has proposed efficient and scalable GP decentralized data fusion algorithms for cooperative perception of environmental phenomena that can distribute the computational load among the mobile sensing agents. The intuition of the GP-DDF algorithm of Chen et al. (2015) is as follows: Each of the N mobile sensing agents constructs a local summary of the data/observations taken along its own path based on a common support set $\mathcal{S} \subset \mathcal{X}$ known to all the other agents and communicates its local summary to them. Then, it assimilates the local summaries received from the other agents into a globally consistent summary which is used to compute a Gaussian predictive distribution for predicting the phenomenon. Formally, the local and global summaries and the Gaussian predictive distribution induced by GP-DDF are defined as follows:

Definition 1 (Local Summary) *Given a common support set $\mathcal{S} \subset \mathcal{X}$ known to all N mobile sensing agents, each agent i encapsulates a column vector $y_{\mathcal{D}_i}$ of realized measurements for its observed locations $\mathcal{D}_i$ into a local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ where*

$$\begin{array}{rcl} \nu_{\mathcal{B}|\mathcal{D}_i} & \triangleq & \Sigma_{\mathcal{B}\mathcal{D}_i}\Sigma_{\mathcal{D}_i\mathcal{D}_i|\mathcal{S}}^{-1}(y_{\mathcal{D}_i} - \mu_{\mathcal{D}_i}) \, , \\ \Psi_{\mathcal{B}\mathcal{B}'|\mathcal{D}_i} & \triangleq & \Sigma_{\mathcal{B}\mathcal{D}_i}\Sigma_{\mathcal{D}_i\mathcal{D}_i|\mathcal{S}}^{-1}\Sigma_{\mathcal{D}_i\mathcal{B}'} \end{array} \quad (3)$$

for all $\mathcal{B}, \mathcal{B}' \subset \mathcal{X}$ and $\Sigma_{\mathcal{D}_i\mathcal{D}_i|\mathcal{S}} \triangleq \Sigma_{\mathcal{D}_i\mathcal{D}_i} - \Sigma_{\mathcal{D}_i\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{D}_i}$.

Definition 2 (Global Summary) *Given a common support set $\mathcal{S} \subset \mathcal{X}$ known to all N mobile sensing agents and the local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ of every agent $i = 1, \ldots, N$, a global summary is defined as a tuple $(\dot{\nu}_{\mathcal{S}}, \dot{\Psi}_{\mathcal{S}\mathcal{S}})$ where*

$$\dot{\nu}_{\mathcal{S}} \triangleq \textstyle\sum_{i=1}^{N} \nu_{\mathcal{S}|\mathcal{D}_i} \, , \; \dot{\Psi}_{\mathcal{S}\mathcal{S}} \triangleq \sum_{i=1}^{N} \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i} + \Sigma_{\mathcal{S}\mathcal{S}} \, . \quad (4)$$

Definition 3 (GP-DDF) *Given a common support set $\mathcal{S} \subset \mathcal{X}$ known to all N agents and the global summary $(\dot{\nu}_{\mathcal{S}}, \dot{\Psi}_{\mathcal{S}\mathcal{S}})$, the GP-DDF algorithm run by each agent computes a Gaussian predictive distribution $\mathcal{N}(\overline{\mu}_x, \overline{\sigma}_x^2)$ of the measurement for any unobserved location $x \in \mathcal{X} \setminus \mathcal{D}$ where*

$$\begin{array}{rcl} \overline{\mu}_x & \triangleq & \mu_x + \Sigma_{x\mathcal{S}}\dot{\Psi}_{\mathcal{S}\mathcal{S}}^{-1}\dot{\nu}_{\mathcal{S}} \, , \\ \overline{\sigma}_x^2 & \triangleq & \sigma_{xx} - \Sigma_{x\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}}^{-1} - \dot{\Psi}_{\mathcal{S}\mathcal{S}}^{-1})\Sigma_{\mathcal{S}x} \, . \end{array} \quad (5)$$

The Gaussian predictive distribution (5) computed by the GP-DDF algorithm is theoretically guaranteed by Chen et al. (2015) to be equivalent to that induced by the centralized *partially independent training conditional* (PITC) approximation (Quiñonero-Candela and Rasmussen 2005) of the GP model. Running GP-DDF on each of the N agents can, however, reduce the $\mathcal{O}(|\mathcal{D}|((|\mathcal{D}|/N)^2 + |\mathcal{S}|^2))$ time incurred by PITC to only $\mathcal{O}((|\mathcal{D}|/N)^3 + |\mathcal{S}|^3 + |\mathcal{S}|^2 N)$ time, hence scaling considerably better with increasing data size $|\mathcal{D}|$.

Though GP-DDF scales well with big data, it can predict poorly due to information loss caused by summarizing the measurements and correlation structure of the data/observations and sparse coverage of the areas with highly varying measurements by the support set. To address its shortcoming, the GP-DDF$^+$ algorithm of Chen et al. (2015) additionally exploits the data local to an agent to improve the predictions for unobserved locations "close" to its data (in the correlation sense) while preserving the efficiency of GP-DDF by adopting its idea of summarizing information into local and global summaries (Definitions 1 and 2). The Gaussian predictive distribution computed by GP-DDF$^+$ (Ouyang and Low 2017) is theoretically guaranteed by Chen et al. (2015) to be equivalent to that induced by the centralized *partially independent conditional* (PIC) approximation (Snelson and Ghahramani 2007) of the GP model. GP-DDF$^+$ shares the same improvement in scalability over PIC as that of GP-DDF over PITC.

3 GP-DDF with Agent-Centric Support Sets

Transfer Learning. It can be observed from Section 2 that the GP-DDF and GP-DDF$^+$ algorithms depend on a common support set $\mathcal{S}$ known to all N mobile sensing agents, which raises three non-trivial issues previously discussed in Section 1: (a) Their cubic time cost in $|\mathcal{S}|$ prohibits increasing the size of $\mathcal{S}$ too much to preserve their efficiency, which consequently limits the expansion of the domain of the phenomenon for which it can still be covered and predicted well; (b) if $\mathcal{S}$ sparsely covers the large-scale phenomenon due to its restricted size and is thus "far" from the data and unobserved locations to be predicted, then the values of the components in terms like $\Sigma_{\mathcal{S}\mathcal{D}_i}$ and $\Sigma_{x\mathcal{S}}$ tend to zero, which degrade their predictive performance; and (c) when switching to a new support set, they have to wastefully discard all previous summaries based on the old support set.

To address the above issues, a straightforward approach inspired by the local GPs method is to partition the domain of the phenomenon into local areas and run GP-DDF or GP-DDF$^+$ with a different, sufficiently dense support set for each local area. Such an approach often suffers from discontinuities in predictions and very high predictive uncertainty at the boundaries between local areas (see Fig. 1b) and only utilizes the data within a local area for its predictions, thereby performing poorly in local areas with little/no data. These drawbacks motivate the need to design and develop a transfer learning mechanism for a team of mobile sensing agents capable of sharing and transferring information encapsulated in a summary based on a support set for a local area to that utilizing a different support set for another area. In this section, we will describe our novel transfer learning

mechanism and its use in our GP-DDF or GP-DDF$^+$ algorithm with agent-centric support sets and theoretically bound and analyze its resulting loss of information.

Specifically, supposing a mobile sensing agent i moves from a local area with support set $\mathcal{S}$ to another local area with a different support set $\mathcal{S}'$ (i.e., $\mathcal{S}\bigcap\mathcal{S}' = \emptyset$), the local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ based on the new support set $\mathcal{S}'$ can be derived *exactly* from the local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ utilizing the old support set $\mathcal{S}$ only when the data $(\mathcal{D}_i, y_{\mathcal{D}_i})$ gathered by agent i (i.e., discarded after encapsulating into $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$) in the local area with support set $\mathcal{S}$ can be *fully* recovered from $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$, which is unfortunately not possible. Our key idea is thus to derive the local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ *approximately* from $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ in an efficient and scalable manner by exploiting the following important definition:

Definition 4 (Prior Summary) *Given a support set $\mathcal{S}\subset\mathcal{X}$ for a local area, each mobile sensing agent i encapsulates a column vector $y_{\mathcal{D}_i}$ of realized measurements for its observed locations $\mathcal{D}_i$ into a prior summary $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ where*

$$\begin{array}{rcl}\omega_{\mathcal{S}|\mathcal{D}_i} & \triangleq & \Sigma_{\mathcal{S}\mathcal{D}_i}\Sigma^{-1}_{\mathcal{D}_i\mathcal{D}_i}(y_{\mathcal{D}_i}-\mu_{\mathcal{D}_i})\ ,\\ \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i} & \triangleq & \Sigma_{\mathcal{S}\mathcal{D}_i}\Sigma^{-1}_{\mathcal{D}_i\mathcal{D}_i}\Sigma_{\mathcal{D}_i\mathcal{S}}\ .\end{array} \tag{6}$$

The prior summary $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ (6) is defined in a similar manner to the local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ (3) except for the $\Sigma_{\mathcal{D}_i\mathcal{D}_i}$ term in the former replacing the $\Sigma_{\mathcal{D}_i\mathcal{D}_i|\mathcal{S}}$ term in the latter and is the main ingredient for making our proposed transfer learning mechanism efficient and scalable. Interestingly, the prior summary based on the new support set $\mathcal{S}'$ can be approximated from the prior summary utilizing the old support set S as follows:

Proposition 1 *If $Y_{\mathcal{S}'}$ and $Y_{\mathcal{D}_i}$ are conditionally independent given $Y_{\mathcal{S}}$ (i.e., $\Sigma_{\mathcal{S}'\mathcal{D}_i|\mathcal{S}} = \Sigma_{\mathcal{S}'\mathcal{D}_i} - \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{D}_i} = \underline{0}$) for $i = 1,\ldots,N$, then*

$$\begin{array}{rcl}\omega_{\mathcal{S}'|\mathcal{D}_i} & = & \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\omega_{\mathcal{S}|\mathcal{D}_i},\\ \Phi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i} & = & \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}'}.\end{array} \tag{7}$$

Its proof is in (Ouyang and Low 2017).

Remark. The conditional independence assumption in Proposition 1 extends that on the training conditionals of PITC and PIC (Section 2) which have already assumed conditional independence of $Y_{\mathcal{D}_1},\ldots,Y_{\mathcal{D}_N}$ given $Y_{\mathcal{S}}$. Alternatively, it can be interpreted as a low-rank covariance matrix approximation $\Sigma_{\mathcal{S}'\mathcal{S}}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{D}_i}$ of $\Sigma_{\mathcal{S}'\mathcal{D}_i}$. The quality of this approximation will be theoretically guaranteed later.

To efficiently and scalably derive the local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ approximately from $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$, our transfer learning mechanism will first have to transform the local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ to the prior summary $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ based on the old support set $\mathcal{S}$, then use the latter to approximate the prior summary $(\omega_{\mathcal{S}'|\mathcal{D}_i}, \Phi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ based on the new support set $\mathcal{S}'$ by exploiting Proposition 1, and finally transform the approximated prior summary back to approximate the local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$, as detailed in Algorithm 1 below. The above two transformations can be achieved by establishing the following relationship between the local summary and prior summary:

Proposition 2 *Given a support set $\mathcal{S}\subset\mathcal{X}$ for a local area, the local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ (3) and the prior summary $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ (6) of agent i are related by*

$$\Phi^{-1}_{\mathcal{S}\mathcal{S}|\mathcal{D}_i}\omega_{\mathcal{S}|\mathcal{D}_i} = \Psi^{-1}_{\mathcal{S}\mathcal{S}|\mathcal{D}_i}\nu_{\mathcal{S}|\mathcal{D}_i}\ ,\ \Phi^{-1}_{\mathcal{S}\mathcal{S}|\mathcal{D}_i} = \Psi^{-1}_{\mathcal{S}\mathcal{S}|\mathcal{D}_i} + \Sigma^{-1}_{\mathcal{S}\mathcal{S}}\ . \tag{8}$$

Its proof is in (Ouyang and Low 2017).

Supposing agent i has gathered additional data $(\mathcal{D}'_i, y_{\mathcal{D}'_i})$ from the local area with the new support set $\mathcal{S}'$, it can be encapsulated into a local summary $(\nu_{\mathcal{S}'|\mathcal{D}'_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}'_i})$ that is assimilated with the approximated local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ by simply summing them up:

$$\begin{array}{rcl}\nu_{\mathcal{S}'|\mathcal{D}_i\bigcup\mathcal{D}'_i} & = & \nu_{\mathcal{S}'|\mathcal{D}_i} + \nu_{\mathcal{S}'|\mathcal{D}'_i}\ ,\\ \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i\bigcup\mathcal{D}'_i} & = & \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i} + \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}'_i}\ ,\end{array} \tag{9}$$

which require making a further assumption of conditional independence between $\mathcal{D}'_i$ and $\mathcal{D}_j$ given the support set $\mathcal{S}'$ for $j = 1,\ldots,N$.

Finally, to assimilate the local summary of agent i with the other agents' local summaries (i.e., based on their current choices of support sets) into a global summary to be used for predicting the phenomenon, the local summary $(\nu_{\mathcal{S}'|\mathcal{D}_j}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_j})$ of every other agent $j\neq i$ based on agent i's support set $\mathcal{S}'$ can be derived approximately from the received local summary $(\nu_{\mathcal{S}''|\mathcal{D}_j}, \Psi_{\mathcal{S}''\mathcal{S}''|\mathcal{D}_j})$ based on agent j's support set $\mathcal{S}''\neq\mathcal{S}'$ using exactly the same transfer learning mechanism described above. Then, the global summary $(\dot{\nu}_{\mathcal{S}'}, \dot{\Psi}_{\mathcal{S}'\mathcal{S}'})$ can be computed via (4) and used by the GP-DDF or GP-DDF$^+$ algorithm (Section 2).

Supposing $|\mathcal{S}| = |\mathcal{S}'| = |\mathcal{S}''|$ for simplicity, our transfer learning mechanism in Algorithm 1 incurs only $\mathcal{O}(|\mathcal{S}|^3)$ time (i.e., independent of data size $|\mathcal{D}|$) due to multiplication and inversion of matrices of size $|\mathcal{S}|$ by $|\mathcal{S}|$. Since the support set for every local area is expected to be small, our transfer learning mechanism is efficient and scalable.

Information Loss from Low-Rank Approximation. Recall from the remark after Proposition 1 that our transfer learning mechanism has utilized a low-rank covariance matrix approximation $\Sigma_{\mathcal{S}'\mathcal{S}}\Sigma^{-1}_{\mathcal{S}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{D}_i}$ of $\Sigma_{\mathcal{S}'\mathcal{D}_i}$. To theoretically bound the information loss resulting from such an approximation, we first observe that it resembles the Nyström low-rank approximation except that the latter typically involves approximating a symmetric positive semi-definite matrix like $\Sigma_{\mathcal{S}'\mathcal{S}'}$ or $\Sigma_{\mathcal{D}_i\mathcal{D}_i}$ instead of $\Sigma_{\mathcal{S}'\mathcal{D}_i}$, which precludes a direct application of existing results on Nyström approximation to our theoretical analysis. Fortunately, we can exploit the idea of clustering with respect to $\mathcal{S}$ for our theoretical analysis which is inspired by that of the Nyström approximation of Zhang, Tsang, and Kwok (2008) but results in a different loss bound depending on the GP hyperparameters (Section 2) and the "closeness" of $\mathcal{S}'$ and $\mathcal{D}_i$ to $\mathcal{S}$ in the correlation sense.

Define $c(x)$ as a function mapping each $x\in\mathcal{D}_i\bigcup\mathcal{S}'$ to the "closest" $c(x)\in\mathcal{S}$, that is, $c:\mathcal{D}_i\bigcup\mathcal{S}'\rightarrow\mathcal{S}$ where $c(x)\triangleq\arg\min_{s\in\mathcal{S}}||\Lambda^{-1}(x-s)||$. Then, partition $\mathcal{D}_i$ ($\mathcal{S}'$) into $|\mathcal{S}|$ disjoint subsets $\mathcal{D}_{is}\triangleq\{x\in\mathcal{D}_i\mid c(x)=s\}$ ($\mathcal{S}'_s\triangleq\{x\in\mathcal{S}'\mid c(x)=s\}$) for $s\in\mathcal{S}$. Intuitively, $\mathcal{D}_{is}$ ($\mathcal{S}'_s$) is

Algorithm 1: GP-DDF/GP-DDF$^+$ with agent-centric support sets based on transfer learning for agent i

if agent i transits from local area with support set $\mathcal{S}$ to local area with support set $\mathcal{S}'$ **then**
- `/* Transfer learning mechanism */`
- Construct local summary $(\nu_{\mathcal{S}|\mathcal{D}_i}, \Psi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ and transform it to prior summary $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ by (8);
- Derive prior summary $(\omega_{\mathcal{S}'|\mathcal{D}_i}, \Phi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ based on $\mathcal{S}'$ approximately from $(\omega_{\mathcal{S}|\mathcal{D}_i}, \Phi_{\mathcal{S}\mathcal{S}|\mathcal{D}_i})$ by (7);
- Transform prior summary $(\omega_{\mathcal{S}'|\mathcal{D}_i}, \Phi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ to local summary $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ by (8);

if agent i has to predict the phenomenon **then**
- **if** data $(\mathcal{D}'_i, y_{\mathcal{D}'_i})$ is available from local area with support set $\mathcal{S}'$ **then**
  - Assimilate local summaries $(\nu_{\mathcal{S}'|\mathcal{D}_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i})$ with $(\nu_{\mathcal{S}'|\mathcal{D}'_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}'_i})$ to yield $(\nu_{\mathcal{S}'|\mathcal{D}_i \cup \mathcal{D}'_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i \cup \mathcal{D}'_i})$ by (9);
- Exchange local summary with every agent $j \neq i$;
- **foreach** agent $j \neq i$ in local area with support set $\mathcal{S}'' \neq \mathcal{S}'$ **do**
  - Derive local summary $(\nu_{\mathcal{S}'|\mathcal{D}_j}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_j})$ based on $\mathcal{S}'$ approximately from received local summary $(\nu_{\mathcal{S}''|\mathcal{D}_j}, \Psi_{\mathcal{S}''\mathcal{S}''|\mathcal{D}_j})$ based on $\mathcal{S}''$ using the above transfer learning mechanism;
- Compute global summary $(\dot{\nu}_{\mathcal{S}'}, \dot{\Psi}_{\mathcal{S}'\mathcal{S}'})$ by (4) using local summaries $(\nu_{\mathcal{S}'|\mathcal{D}_i \cup \mathcal{D}'_i}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_i \cup \mathcal{D}'_i})$ and $(\nu_{\mathcal{S}'|\mathcal{D}_j}, \Psi_{\mathcal{S}'\mathcal{S}'|\mathcal{D}_j})$ of every agent $j \neq i$;
- Run GP-DDF or GP-DDF$^+$ (Section 2);

a cluster of locations in $\mathcal{D}_i$ ($\mathcal{S}'$) that are closest to location s in the support set $\mathcal{S}$. Our main result below theoretically bounds the information loss $||\Sigma_{\mathcal{S}'\mathcal{D}_i} - \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{D}_i}||_F$ resulting from the low-rank approximation $\Sigma_{\mathcal{S}'\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{D}_i}$ of $\Sigma_{\mathcal{S}'\mathcal{D}_i}$ with respect to the Frobenius norm:

Theorem 1 *Let $\sigma_{xx'}$ be defined by a squared exponential covariance function* (1)*, $T \triangleq \arg\max_{s\in\mathcal{S}} |\mathcal{D}_{is}|$, $T' \triangleq \arg\max_{s\in\mathcal{S}} |\mathcal{S}'_s|$, $\epsilon_{\mathcal{S}'} \triangleq |\mathcal{S}'|^{-1}\sum_{x\in\mathcal{S}'} ||\Lambda^{-1}(x - c(x))||^2$, and $\epsilon_{\mathcal{D}_i} \triangleq |\mathcal{D}_i|^{-1}\sum_{x\in\mathcal{D}_i} ||\Lambda^{-1}(x - c(x))||^2$. Then,*

$$||\Sigma_{\mathcal{S}'\mathcal{D}_i} - \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{D}_i}||_F \leq \sqrt{3/e}\sigma_s^2|\mathcal{S}|TT'(\sqrt{\epsilon_{\mathcal{S}'}} + \sqrt{\epsilon_{\mathcal{S}'} + \epsilon_{\mathcal{D}_i}} + \sqrt{\epsilon_{\mathcal{D}_i}} + \sigma_s^2||\Sigma_{\mathcal{S}\mathcal{S}}^{-1}||_F|\mathcal{S}|\sqrt{3\epsilon_{\mathcal{S}'}\epsilon_{\mathcal{D}_i}/e}) .$$

Its proof is in (Ouyang and Low 2017). Note that a similar result to Theorem 1 can be derived for other commonly-used covariance functions such as those presented in the work of Zhang, Tsang, and Kwok (2008). It can be observed from Theorem 1 that the information loss $||\Sigma_{\mathcal{S}'\mathcal{D}_i} - \Sigma_{\mathcal{S}'\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{D}_i}||_F$ can be reduced when the signal variance σ_s^2 is small, the length-scales ℓ_1 and/or ℓ_2 are large, the mobile sensing agent i utilizes a support set $\mathcal{S}$ "close" to its observed locations $\mathcal{D}_i$ in a local area (i.e., smaller $\epsilon_{\mathcal{D}_i}$) and moves to another local area with a support set $\mathcal{S}'$ "close" to $\mathcal{S}$ (i.e., smaller $\epsilon_{\mathcal{S}'}$).

Lazy Transfer Learning. Theorem 1 above further reveals that every instance of transfer learning in Algorithm 1 incurs some information loss which accumulates over multiple instances when the agent transits between many local areas and consequently degrades its resulting predictive performance. This motivates the need to be frugal in the number of instances of transfer learning to be performed.

To achieve this, our key idea is to delay transfer learning till prediction time but in a memory-efficient manner[1]. Specifically, we propose the following new information sharing mechanism to reduce memory requirements for a team of mobile sensing agents: When agent i leaves a local area, its local summary is communicated to another agent in the same area who assimilates it with its own local summary using (4). However, if no other agent is in the same area, then agent i stores a backup of its local summary. On the other hand, when agent i enters a local area containing other agents, it simply obtains its corresponding support set to encapsulate its new data gathered in this area. But, if no other agent is in this area, then agent i retrieves (and removes) the backup of its corresponding local summary from an agent who has previously visited this area[2]. If no agent has such a backup, then agent i is the first to visit this area and constructs a new support set for it. Algorithm 2 in (Ouyang and Low 2017) details GP-DDF/GP-DDF$^+$ with agent-centric support sets by incorporating the above information sharing mechanism in order to achieve memory-efficient lazy transfer learning.

To analyze the memory requirements of our information sharing mechanism in Algorithm 2 in (Ouyang and Low 2017), let the domain of the phenomenon be partitioned into K local areas. Then, the team of N mobile sensing agents incurs a total of $\mathcal{O}((K + N)|\mathcal{S}|^2)$ memory in the worst case when all the agents reside in the same local area and the last agent entering this area stores the backups of the local summaries for the other $K - 1$ local areas. However, the agents are usually well-distributed over the entire phenomenon in practice: In the case of evenly distributed agents, the team incurs a total of $\mathcal{O}(\max(K, N)|\mathcal{S}|^2)$ memory. So, each agent incurs an amortized memory cost of $\mathcal{O}(\max(K, N)|\mathcal{S}|^2/N)$.

A limitation of the information sharing mechanism in Algorithm 2 in (Ouyang and Low 2017) is its susceptibility to agent failure: If an agent stores the backups of the local summaries for many local areas and breaks down, then all the information on these local areas will be lost. Its robustness to agent failure can be improved by distributing multiple agents to every local area to reduce its risk of being empty and hence its likelihood of inducing a backup.

4 Experiments and Discussion

This section empirically evaluates the performance of our GP-DDF and GP-DDF$^+$ algorithms with agent-centric sup-

[1]Naively, an agent can delay transfer learning by simply storing a separate local summary based on the support set for every previously visited local area, which is not memory-efficient.

[2]Multiple backups of the local summary for the same local area may exist if agents leave this area at the same time, which rarely happens. In this case, agent i should retrieve (and remove) all these backups from the agents storing them.

port sets using simulated spatial phenomena (Section 4.1) and two real-world environmental phenomena (Section 4.2).

Performance Metrics. Two performance metrics are used in our experiments: (a) *Root-mean-square error* (RMSE) $\sqrt{|\mathcal{X}|^{-1}\sum_{x\in\mathcal{X}}(\overline{\mu}_x - y_x)^2}$ measures the predictive performance of the tested algorithms while (b) incurred time measures their efficiency and scalability.

4.1 Simulated Spatial Phenomena

The simulated experiment here is set up to demonstrate the effectiveness of our proposed lazy transfer learning mechanism (Section 3) that is driving our GP-DDF/GP-DDF$^+$ algorithms with agent-centric support sets (Ouyang and Low 2017): A number of 2-dimensional spatial phenomena of size 50 by 50 are generated using signal variance $\sigma_s^2 = 1$, noise variance $\sigma_n^2 = 0.01$, and by varying the length-scale $\ell_1 = \ell_2$ from 1 to 20. The domain of the spatial phenomenon is partitioned into 4 disjoint local areas of size 25 by 25 (Fig. 1), each of which contains an agent moving randomly within to gather 25 local data/observations. We compare the predictive performance of the following decentralized data fusion algorithms: (a) Original GP-DDF (Chen et al. 2012; 2015) with a common support set of size 18 uniformly distributed over the entire phenomenon and known to all 4 agents, (b) *PITCs utilizing local information* (local PITCs) with agent-centric support sets assign a different PITC to each agent summarizing its gathered local data based on a support set of size 18 uniformly distributed over its residing local area, (c) *PITCs utilizing full information* (full PITCs) with agent-centric support sets assign a different PITC to each agent summarizing its gathered local data as well as those communicated by the other agents (i.e., full data gathered by all agents) based on a support set of size 18 uniformly distributed over its residing local area, (d) *GP-DDF with agent-centric support sets* (GP-DDF-ASS) each of size 18 and uniformly distributed[3] over a different local area (Algorithm 2 in (Ouyang and Low 2017)). Note that if our proposed lazy transfer learning mechanism in GP-DDF-ASS incurs minimal (total) information loss, then its predictive performance will be similar to that of full PITCs (local PITCs).

Fig. 1 shows results of the maps of log-predictive variance (i.e., $\log\overline{\sigma}_x^2$ for all $x \in \mathcal{X}$) over a spatial phenomenon with length-scale of 10 achieved by the tested decentralized data fusion algorithms. It can be observed from Fig. 1a that GP-DDF achieves the worst predictive performance since its common support set, which is uniformly distributed over the entire phenomenon, is of the same size as an agent-centric support set uniformly distributed over each of the 4 smaller disjoint local areas to be used by the other tested algorithms. From Fig. 1b, though local PITCs can predict better than GP-DDF, the predictive uncertainty at the boundaries between local areas remains very high, which is previously explained in Section 3. Fig. 1c shows the most ideal predictive performance achieved by full PITCs because each agent exploits the full data gathered by and exchanged with all agents for encapsulating into a global summary based on the support set distributed over its residing local area. Fig. 1d reveals that GP-DDF-ASS can achieve predictive performance comparable to that of full PITCs without needing to exchange the full data between all agents due to minimal information loss by our lazy transfer learning mechanism.

Recall from Theorem 1 (Section 3) that the information loss incurred by our proposed transfer learning mechanism depends on the closeness between the support sets distributed over different local areas as well as the closeness (i.e., in the correlation sense) between the support sets and the data/observations. The effect of varying such closeness on the performance of our transfer learning mechanism can be empirically investigated by alternatively changing the length-scale to control the degree of spatial correlation between the measurements of the phenomenon. Fig. 1e shows results of the reduction in RMSE of GP-DDF, full PITCs, and GP-DDF-ASS over local PITCs with varying lengthscales from 1 to 20. It can be observed that only GP-DDF performs worse than local PITCs while both GP-DDF-ASS and full PITCs perform significantly better than local PITCs, all of which are explained previously. Interestingly, the reduction in RMSEs varies for different length-scales and tends to zero when the length-scale is either too small or large. With a very small length-scale, the correlations between the support sets distributed over different local areas and between the support sets and the data/observations become near-zero, hence resulting in poor transfer learning for GP-DDF-ASS. This agrees with the observation in our theoretical analysis for Theorem 1 (Section 3). With a very large length-scale, though their correlations are strong, the local observations/data can be used by local PITCs to predict very well, hence making transfer learning redundant. Our transfer learning mechanism performs best with intermediate length-scales where the correlations between the support sets distributed over different local areas and between the support sets and the data are sufficiently strong but not to the extent of achieving good predictions with simply local data.

4.2 Real-World Environmental Phenomena

The performance of our GP-DDF and GP-DDF$^+$ algorithms with agent-centric support sets are empirically evaluated using the following two real-world datasets (as well as the MODIS plankton density dataset in (Ouyang and Low 2017)): (a) The indoor lighting quality dataset contains 1200 observations of relative lighting level gathered simultaneously by three real Pioneer 3-DX mobile robots mounted with SICK LMS200 laser rangefinders and weather boards while patrolling an office environment, as shown in (Ouyang and Low 2017). The domain of interest is partitioned into $K = 8$ consecutive local areas and the robots patrol to and fro across them such that they visit all $K = 8$ local areas exactly twice to gather observations of relative lighting level; and (b) the monthly sea surface temperature ($^\circ$C) dataset (Ouyang and Low 2017) is bounded within lat. 35.75-14.25S and lon. 80.25-104.25E (i.e., in the In-

[3]Alternatively, active learning can be used to select an informative support set *a priori* for each local area (Chen et al. 2015). Empirically, this yields little performance improvement due to a sufficiently dense (yet small) support set uniformly distributed over the local area and slightly beyond its boundary by 10% of its width.

dian ocean) and gathered from Dec. 2002 to Dec. 2015 with a data size of 1083608. The huge spatiotemporal domain of this phenomenon comprises 5-dimensional input feature vectors of latitude, longitude, year, month, and season, and is spatially partitioned into 32 disjoint local areas, each of which is temporally split into 64 disjoint intervals (hence, $K = 2048$) and assigned 2 agents moving randomly within to gather local observations (hence, a total of 64 agents); the results are averaged over 10 runs.

The performance of our *GP-DDF and GP-DDF$^+$ algorithms with agent-centric support sets* (respectively, GP-DDF-ASS and GP-DDF$^+$-ASS), each of which is of size 64 (324) and uniformly distributed[3] over a different local area of the office environment (temperature phenomenon), are compared against that of the local GPs[4] method and state-of-the-art GP-DDF and GP-DDF$^+$ (Chen et al. 2015) with a common support set of size 64 (324) uniformly distributed over the entire office environment (temperature phenomenon) and known to all agents; consequently, the latter construct local summaries of the same size. The hyperparameters of GP-DDF-ASS and GP-DDF$^+$-ASS are learned using maximum likelihood estimation, as detailed in (Ouyang and Low 2017).

Predictive Performance. Figs. 2a and 2c show results of decreasing RMSE achieved by tested algorithms with an increasing total number of observations, which is expected. It can be observed that GP-DDF-ASS and GP-DDF$^+$-ASS, respectively, outperform GP-DDF and GP-DDF$^+$, as explained previously in the last paragraph of Section 1. Furthermore, the performance improvement of GP-DDF-ASS over GP-DDF is larger than that of GP-DDF$^+$-ASS over GP-DDF$^+$, which demonstrates the effectiveness of our lazy transfer learning mechanism, especially when some local areas lack data/observations. This also explains the better predictive performance of GP-DDF$^+$-ASS over local GPs, even though they both exploit local data.

Time Efficiency. In this experiment, we specifically evaluate the time efficiency of our transfer learning mechanism (Section 3) in GP-DDF-ASS and GP-DDF$^+$-ASS with respect to the number of observations; to do this, we have intentionally ignored the time incurred by their information sharing mechanism (i.e., first if-then construct in Algorithm 2 in (Ouyang and Low 2017)) and compared their resulting incurred time with that of GP-DDF and GP-DDF$^+$ (i.e., without transfer learning). Figs. 2b and 2d show results of increasing total time incurred by tested algorithms when the total number of observations increases, which is expected (Section 2). It can be observed that GP-DDF-ASS and GP-DDF$^+$-ASS, respectively, incur only slightly more time than GP-DDF and GP-DDF$^+$ (i.e., due to an extra small fixed cost of $\mathcal{O}(|\mathcal{S}|^3)$ time for transfer learning (Section 3)) to achieve more superior predictive performance, especially for GP-DDF-ASS. GP-DDF$^+$-ASS incurs more time than GP-DDF-ASS (local GPs) to further exploit local data (support set and transfer learning) for improving its predictive performance. For *time-critical applications*, we recommend using GP-DDF-ASS over GP-DDF$^+$-ASS since its incurred time is small and increases very gradually with more observations while its performance improvement over GP-DDF is significant. For *big data applications*, GP-DDF$^+$-ASS is instead preferred since a large amount of local data is often available in nearly every local area for prediction.

Scalability in the Number of Agents. Fig. 2e shows results of total time incurred by tested algorithms averaged over 30 runs with an increasing number N of agents (i.e., up to 128 agents) to gather a total number of 1235 observations from a plankton density phenomenon; the experimental setup is detailed in (Ouyang and Low 2017). It can be observed that the total time incurred by GP-DDF-ASS and GP-DDF$^+$-ASS decrease with more agents, as explained in Section 2, and they, respectively, incur only slightly more time than GP-DDF and GP-DDF$^+$ due to their information sharing mechanism described in Section 3 (i.e., first if-then construct in Algorithm 2 in (Ouyang and Low 2017)). Additional empirical results and analysis for the plankton density phenomenon are reported in (Ouyang and Low 2017).

[4]Local GPs result from a sparse block-diagonal $\Sigma_{\mathcal{D}\mathcal{D}}$ (2).

5 Conclusion

This paper describes novel GP-DDF-ASS and GP-DDF$^+$-ASS algorithms for distributed cooperative perception of large-scale environmental phenomena. To overcome the limitations of scale of GP-DDF and GP-DDF$^+$, our proposed algorithms employ a novel transfer learning mechanism between agents which is capable of sharing and transferring information encapsulated in a summary based on a support set to that utilizing a different support set with some loss that can be theoretically bounded and analyzed. To alleviate the issue of information loss accumulating over multiple instances of transfer learning, GP-DDF-ASS and GP-DDF$^+$-ASS exploit a new information sharing mechanism to achieve memory-efficient lazy transfer learning. Empirical evaluation on real-world datasets show that our transfer learning and information sharing mechanisms make GP-DDF-ASS and GP-DDF$^+$-ASS incur only slightly more time than GP-DDF and GP-DDF$^+$ (i.e., without transfer learning) to achieve more superior predictive performance.

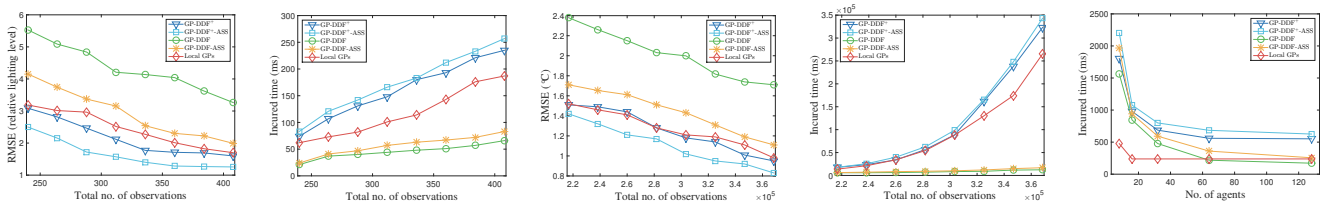
Acknowledgments. This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2, MOE2016-T2-2-156.

References

Cao, N.; Low, K. H.; and Dolan, J. M. 2013. Multi-robot informative path planning for active sensing of environmental phenomena: A tale of two algorithms. In *Proc. AAMAS*.

Chen, J.; Low, K. H.; Tan, C. K.-Y.; Oran, A.; Jaillet, P.; Dolan, J. M.; and Sukhatme, G. S. 2012. Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena. In *Proc. UAI*, 163–173.

Chen, J.; Cao, N.; Low, K. H.; Ouyang, R.; Tan, C. K.-Y.; and Jaillet, P. 2013. Parallel Gaussian process regression with low-rank covariance matrix approximations. In *Proc. UAI*, 152–161.



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

Figure 2: Graphs of RMSE and total time incurred by tested algorithms vs. total no. of observations for (a-b) indoor lighting quality and (c-d) temperature phenomenon, and (e) graphs of total incurred time vs. no. of agents achieved by tested algorithms for plankton density phenomenon.

Chen, J.; Low, K. H.; Jaillet, P.; and Yao, Y. 2015. Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE Trans. Autom. Sci. Eng.* 12:901–921.

Chen, J.; Low, K. H.; and Tan, C. K.-Y. 2013. Gaussian process-based decentralized data fusion and active sensing for mobility-on-demand system. In *Proc. Robotics: Science and Systems Conference*.

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

Daxberger, E., and Low, K. H. 2017. Distributed batch Gaussian process optimization. In *Proc. ICML*, 951–960.

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.

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*, 1–10.

Hoang, T. N.; Low, K. H.; Jaillet, P.; and Kankanhalli, M. 2014. Nonmyopic ϵ-Bayes-optimal active learning of Gaussian processes. In *Proc. ICML*, 739–747.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2015. A unifying framework of anytime sparse Gaussian process regression models with stochastic variational inference for big data. In *Proc. ICML*, 569–578.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2016. A distributed variational inference framework for unifying parallel sparse Gaussian process regression models. In *Proc. ICML*, 382–391.

Hoang, Q. M.; Hoang, T. N.; and Low, K. H. 2017. A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression. In *Proc. AAAI*, 2007–2014.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2018. Decentralized high-dimensional Bayesian optimization with factor graphs. In *Proc. AAAI*.

Ling, C. K.; Low, K. H.; and Jaillet, P. 2016. Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond. In *Proc. AAAI*, 1860–1866.

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*, 755–760.

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*, 105–112.

Low, K. H.; Yu, J.; Chen, J.; and Jaillet, P. 2015. Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation. In *Proc. AAAI*.

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

Low, K. H.; Dolan, J. M.; and Khosla, P. 2009. Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*.

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

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

Ouyang, R., and Low, K. H. 2017. Gaussian process decentralized data fusion with agent-centric support sets for large-scale distributed cooperative perception. arXiv:1711.06064.

Ouyang, R.; Low, K. H.; Chen, J.; and Jaillet, P. 2014. Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena. In *Proc. AAMAS*.

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

Snelson, E. L., and Ghahramani, Z. 2007. Local and global sparse Gaussian process approximation. In *Proc. AISTATS*.

Xu, N.; Low, K. H.; Chen, J.; Lim, K. K.; and Özgül, E. B. 2014. GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model. In *Proc. AAAI*, 2585–2592.

Zhang, Y.; Hoang, T. N.; Low, K. H.; and Kankanhalli, M. 2016. Near-optimal active learning of multi-output Gaussian processes. In *Proc. AAAI*.

Zhang, K.; Tsang, I. W.; and Kwok, J. T. 2008. Improved Nyström low-rank approximation and error analysis. In *Proc. ICML*, 1232–1239.