AMBROSia: An Autonomous Model-Based Reactive Observing System*

David Caron, Abhimanyu Das, Amit Dhariwal, Leana Golubchik**,
Ramesh Govindan, David Kempe, Carl Oberg, Abhishek Sharma, Beth Stauffer,
Gaurav Sukhatme, and Bin Zhang

University of Southern California, Los Angeles, CA 90089
**leana@cs.usc.edu

Abstract. Observing systems facilitate scientific studies by instrumenting the real world and collecting corresponding measurements, with the aim of detecting and tracking phenomena of interest. Our AMBROSia project focuses on a class of observing systems which are *embedded* into the environment, consist of *stationary and mobile* sensors, and *react* to collected observations by reconfiguring the system and adapting which observations are collected next. In this paper, we report on recent research directions and corresponding results in the context of AMBROSia.

1 Introduction

Observing systems facilitate scientific studies by instrumenting the real world and collecting measurements, with the aim of detecting and tracking phenomena of interest. Our work focuses Reactive Observing Systems (ROS), i.e., those that are (1) *embedded* into the environment, (2) consist of *stationary and mobile* sensors, and (3) *react* to collected observations by reconfiguring the system and adapting which observations are collected next. The goal of ROS is to help scientists verify or falsify hypotheses with useful samples taken by the stationary and mobile units, as well as to analyze data autonomously to discover interesting trends or alarming conditions. We explore ROS in the context of a marine biology application, where the system monitors, e.g., water temperature and light as well as concentrations of micro-organisms and algae in a body of water.

Current technology (and its realistic near future prediction) precludes sampling all possibly relevant data: bandwidth limitations between stationary sensors make it impossible to collect all sensed data, and time & storage capacity constraints for mobile entities curtail the number and locations of samples they can take. To make good use of limited resources, we are developing a framework capable of optimizing and controlling the set of samples to be taken at any given time, taking into consideration the

* This research has been funded by the NSF DDDAS 0540420 grant. It has also been funded in part by the NSF Center for Embedded Networked Sensing Cooperative Agreement CCR-0120778, the National Oceanic and Atmospheric Administration Grant NA05NOS47812228, and the NSF EIA-0121141 grant.

** Contact author.

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 995–1001, 2007.

application's objectives and system resource constraints. We refer to this framework as AMBROSia (Autonomous Model-Based Reactive Observing System). In [7] we give an overview of the AMBROSia framework as well as the experimental system setting and the corresponding marine biology application. In this paper we report on our recent research directions and corresponding results, in the context of AMBROSia.

As already noted, one of the core functionalities of AMBROSia is the selection of samples which (1) can be retrieved at reasonably low energy cost, and (2) yield as much information as possible about the system. The second property in particular will change dynamically: in reaction to past measurements, different observations may be more or less useful in the future. At any given time, the system must select the most informative samples to retrieve based on the model at that point. We briefly outline a mathematical formulation of this problem and results to date in Section 2.

In ROS accurate measurements are useful to scientists seeking a better understanding of the environment. However, it may not be feasible to move the static sensor nodes after deployment. In such cases, mobile robots could be used to augment the static sensor network, hence forming a robotic sensor network. In such networks, an important question to ask is how to coordinate the mobile robots and the static nodes such that estimation errors are minimized. Our recent efforts on addressing this question are briefly outlined in Section 3.

The successful use of ROS, as envisioned in AMBROSia, partly depends on the system's ability to ensure the collected data's quality. However, various sensor network measurement studies have reported transient faults in sensor readings. Thus, another important goal in AMBROSia is automated high-confidence fault detection, classification, and data rectification. As a first step towards that goal, we explore and characterize several qualitatively different classes of fault detection methods, which are briefly outlined in Section 4.

Our concluding remarks are given in Section 5.

2 A Mathematical Formulation of Sample Selection

Mathematically, our sample selection problem can be modeled naturally as a subset selection problem for regression: Based on the small number of measurements X_i taken, a random variable Z (such as average temperature, chlorophyll concentration, growth of algae, etc.) is to be estimated as accurately as possible. Different measurements X_i, X_j may be partially correlated, and thus partially redundant, a fact that should be deduced from past models. In a pristine and abstract form, the problem can thus be modeled as follows:

We are given a covariance matrix C between the random variables X_i, and a vector $\mathbf{b}$ describing covariances between measurements X_i and the quantity Z to be predicted (C and $\mathbf{b}$ are estimated based on the model). In order to keep the energy sampling cost small, the goal is to find a small set S (of size at most k) so as to minimize the *mean squared prediction error* [4,8] $\mathrm{Err}(Z,S) := E[(Z - \sum_{i \in S} \alpha_i X_i)^2]$, where the α_i are the optimal regression coefficients specifically for the set S selected.

The selection problem thus gives rise to the well-known *subset selection problem for regression* [10], which has traditionally had many applications in medical and social studies, where the set S is interpreted as a good predictor of Z. Finding the best set S of size k is NP-hard, and certain approximation hardness results are known [2,11]. However, despite its tremendous importance to statistical sciences, very little was known in terms of approximation algorithms until recent results by Gilbert et al. [6] and Tropp [17] established approximation guarantees for the very special case of nearly independent X_i variables.

In ongoing work, we are investigating several more general cases of the subset selection problem for regression, in particular with applications to selecting samples to draw in sensor network environments. Over the past year, we have obtained the following key results (which are currently under submission [1]):

Theorem 1. *If the pairwise covariances between the X_i are small (at most $1/6k$, if k variables can be selected), then the frequently used Forward Regression heuristic is a provably good approximation.*

The quality of approximation is characterized precisely in [1], but omitted here due to space constraints. This result improves on the ones of [6,17], in that it analyzes a more commonly used algorithm, and obtains somewhat improved bounds. The next theorem extends the result to a significantly wider class of covariance matrices, where several pairs can have higher covariances.

Theorem 2. *If the pairs of variables X_i with high covariance (exceeding $\Omega(1/4k)$) form a tree, then a provably good approximation can be obtained in polynomial time using rounding and dynamic programming.*

While this result significantly extends the cases that can be approximated, it is not directly relevant to measuring physical phenomena. Hence, we also study the case of sensors embedded in a metric space, where the covariance between sensors' readings is a monotone decreasing function of their distance. The general version of this problem is the subject of ongoing work, but [1] contains a promising initial finding:

Theorem 3. *If the sensors are embedded on a line (in one dimension), and the covariance decreases roughly exponentially in the distance, then a provably good approximation can be obtained in polynomial time.*

The algorithm is again based on rounding and a different dynamic program, and makes use of some remarkable properties of matrix inverses for this type of covariance matrix. At the moment, we are working on extending these results to more general metrics (in particular, two-dimensional Euclidean metrics), and different dependencies of covariances on the distance.

3 Scalar Field Estimation

Sensor networks provide new tools for observing and monitoring the environment. In aquatic environments, accurately measuring quantities such as temperature, chlorophyll, salinity, and concentration of various nutrients is useful to scientists seeking a better understanding of aquatic ecosystems, as well as government officials charged with ensuring public safety via appropriate hazard warning and remediation measures.

Broadly speaking, these quantities of interest are scalar fields. Each is characterized by a single scalar quantity which varies spatiotemporally. Intuitively, the more the readings near the location where a field estimate is desired, the less the reconstruction error. In other words, the spatial distribution of the measurements (the *samples*) affects the estimation error. In many cases, it may not be feasible to move the static sensor nodes after deployment. In such cases, one or more mobile robots could be used to augment the static sensor network, hence forming a sensor-actuator network or a robotic sensor network.

The problem of adaptive sampling: An immediate question to ask is how to coordinate the mobile robots and the static nodes such that the error associated with the estimation on the scalar field is minimized subject to the constraint that the energy available to the mobile robot(s) is bounded. Specifically, if each static node makes a measurement in its vicinity, and the total energy available to the mobile robot is known, what path should the mobile robot take to minimize the mean square integrated error associated with the reconstruction of the entire field? Here we assume that the energy consumed by communications and sensing is negligible compared to the energy consumed in moving the mobile robot. We also assume that the mobile robot can communicate with all the static nodes and acquire sensor readings from them. Finally, we focus on reconstructing phenomena which do not change temporally(or change very slowly compared to the time it takes the mobile robot to complete a tour of the environment).

The domain: We develop a general solution to the above problem and test it on a particular set up designed to monitor an aquatic environment. The experimental set up is a systems of anchored buoys (the static nodes), and a robotic boat (the mobile robot) capable of measuring temperature and chlorophyll concentrations. This testbed is part of the NAMOS (Networked Aquatic Microbial Observing System) project (`http://robotics.usc.edu/~namos`), which is used in studies of microbial communities in freshwater and marine environments [3,15].

Contributions: We propose an adaptive sampling algorithm for a mobile sensor network consisting of a set of static nodes and a mobile robot tasked to reconstruct a scalar field. Our algorithm is based on local linear regression [13,5]. Sensor readings from static nodes (a set of buoys) are sent to the mobile robot (a boat) and used to estimate the Hessian Matrix of the scalar field (the surface temperature of a lake), which is directly related to the estimation error. Based on this information, a path planner generates a path for the boat such that the resulting integrated mean square error (IMSE) of the field reconstruction is minimized subject to the constraint that the boat has a finite amount of energy which it can expend on the traverse. Data from extensive (several km) traverses in the field as well as simulations, validate the performance of our algorithm.

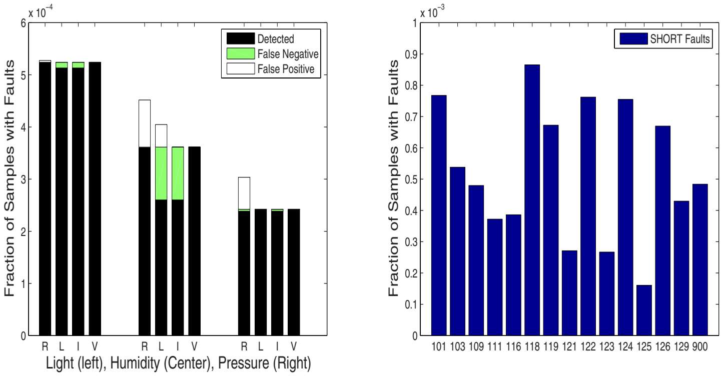
We are currently working on how to determine the appropriate resolution to discretize the sensed field. One interesting observation from the simulations and experiments is that when the initial available energy is increased, the estimation errors decrease rapidly and level off instead of decreasing to zero. Theoretically, when the energy available to the mobile node increases, more sensor readings can be taken and hence the estimation errors should keep decreasing. By examining the path generated by the adaptive sampling algorithm, we found that when the initial energy is enough for the mobile node to go through all the ‘important’ locations, increasing the initial energy does not have much effect on the estimation error. We plan to investigate advanced path planning strategies and alternative sampling design strategies in future work.

4 Faults in Sensor Data

With the maturation of sensor network software, we are increasingly seeing longer-term deployments of wireless sensor networks in real world settings. As a result, research attention is now turning towards drawing meaningful scientific inferences from the collected data [16]. Before sensor networks can become effective replacements for existing scientific instruments, it is important to ensure the quality of the collected data. Already, several deployments have observed faulty sensor readings caused by incorrect hardware design or improper calibration, or by low battery levels [12,16].

Given these observations, and the realization that it will be impossible to always deploy a perfectly calibrated network of sensors, an important research direction for the future will be automated detection, classification, and root-cause analysis of sensor faults, as well as techniques that can automatically scrub collected sensor data to ensure high quality. A first step in this direction is an understanding of the prevalence of faulty sensor readings in existing real-world deployments.

We focus on a small set of sensor faults that have been observed in real deployments: single-sample spikes in sensor readings (SHORT faults), longer duration noisy readings (NOISE faults), and anomalous constant offset readings (CONSTANT faults). Given these fault models, our work makes the following two contributions.



(a) SHORT Faults

(b) SHORT Faults: Light Sensor

Fig. 1. GDI data set

Detection Methods. We have explored three qualitatively different techniques for automatically detecting such faults from a trace of sensor readings. Rule-based methods leverage domain knowledge to develop heuristic rules for detecting and identifying faults. Linear Least-Squares Estimation (LLSE) based methods predict “normal” sensor behavior by leveraging sensor correlation, flagging deviations from the normal as sensor faults. Finally, learning-based methods (based on Hidden Markov Models) are trained to statistically detect and identify classes of faults.

Our findings indicate that these methods sit at different points on the accuracy/robustness spectrum. While rule-based methods can detect and classify faults, they can be sensitive to the choice of parameters. By contrast, the LLSE method is a bit more robust to parameter choices but relies on spatial correlations and cannot classify faults. Finally, our learning method (based on Hidden Markov Models) is cumbersome, partly because it requires training, but it can fairly accurately detect and classify faults. We also explored hybrid detection techniques, which combine these three methods in ways that can be used to reduce false positives or false negatives, whichever is more important for the application. These results are omitted for brevity and the interested reader is referred to [14].

Evaluation on Real-World Datasets. We applied our detection methods to real-world data sets. Here, we present results from the Great Duck Island (GDI) data set [9], where we examine the fraction of faulty samples in a sensor trace.

The predominant fault in the readings was of the type SHORT. We applied the SHORT rule, the LLSE method, and Hybrid(I) (a hybrid detection technique) to detect SHORT faults in light, humidity and pressure sensor readings. Figure 1(a) shows the overall prevalence (computed by aggregating results from all the 15 nodes) of SHORT faults for different sensors in the GDI data set. (On the x-axis of this figure, the SHORT rule's label is **R**, LLSE's label is **L**, and Hybrid(I)'s label is **I**.) The Hybrid (I) technique eliminates any false positives reported by the SHORT rule or the LLSE method. The intensity of SHORT faults was high enough to detect them by visual inspection of the entire sensor readings timeseries. This ground-truth is included for reference in the figure under the label **V**. It is evident from the figure that SHORT faults are relatively infrequent. They are most prevalent in the light sensor readings (approximately 1 fault every 2000 samples). Figure 1(b) shows the distribution of SHORT faults in light sensor readings across various nodes. (Here, node numbers are indicated on the x-axis.) SHORT faults do not exhibit any discernible pattern in the prevalence of these faults across different sensor nodes; the same holds for other sensors, but we have omitted the corresponding graphs for brevity.For results on other data sets, please refer to [14].

Our study informs the research on ensuring data quality. Even though we find that faults are relatively rare, they are not negligibly so, and careful attention needs to be paid to engineering the deployment and to analyzing the data. Furthermore, our detection methods could be used as part of an online fault diagnosis system, i.e., where corrective steps could be taken during the data collection process based on the diagnostic system's results.

5 Concluding Remarks

Overall, our vision for AMBROSia is that it will facilitate observation, detection, and tracking of scientific phenomena that were previously only partially (or not at all) observable and/or understood. In this paper we outlined results corresponding to some of our recent steps towards achieving this vision.

References

1. A. Das and D. Kempe. Algorithms for subset selection in regression, 2006. Submitted to STOC 2007.
2. G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *Journal of Constructive Approximation*, 13:57–98, 1997.
3. Amit Dhariwal, Bin Zhang, Carl Oberg, Beth Stauffer, Aristides Requicha, David Caron, and Gaurav S. Sukhatme. Networked aquatic microbial observing system. In *the Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, May 2006.
4. G. Diekhoff. *Statistics for the Social and Behavioral Sciences*. Wm. C. Brown Publishers, 2002.
5. Jianqing Fan. Local linear regression smoothers and their minimax efficiences. *The Annals of Statistics*, 21(1):196–216, 1993.
6. A. Gilbert, S. Muthukrishnan, and M. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. ACM-SIAM Symposiun on Discrete Algorithms*, 2003.
7. Leana Golubchik, David Caron, Abhimanyu Das, Amit Dhariwal, Ramesh Govindan, David Kempe, Carl Oberg, Abhishek Sharma, Beth Stauffer, Gaurav Sukhatme, and Bin Zhang. A Generic Multi-scale Modeling Framework for Reactive Observing Systems: an Overview. In *Proceedings of the Dynamic Data Driven Application Systems Workshop held with ICCS*, 2006.
8. R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.
9. Alan Mainwaring, Joseph Polastre, Robert Szewczyk, and David Cullerand John Anderson. Wireless Sensor Networks for Habitat Monitoring . In *the ACM International Workshop on Wireless Sensor Networks and Applications. WSNA '02*, 2002.
10. A. Miller. *Subset Selection in Regression*. Chapman and Hall, second edition, 2002.
11. B. Natarajan. Sparse approximation solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.
12. N. Ramanathan, L. Balzano, M. Burt, D. Estrin, E. Kohler, T. Harmon, C. Harvey, J. Jay, S. Rothenberg, and M. Srivastava. Rapid Deployment with Confidence: Calibration and Fault Detection in Environmental Sensor Networks. Technical Report 62, CENS, April 2006.
13. D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.
14. A. Sharma, L. Golubchik, and R. Govindan. On the Prevalence of Sensor Faults in Real World Deployments. Technical Report 07-888, Computer Science, University of Southern California, 2007.
15. Gaurav S. Sukhatme, Amit Dahriwal, Bin Zhang, Carl Oberg, Beth Stauffer, and David Caron. The design and development of a wireless robotic networked aquatic microbial observing system. *Environmental Engineering Science*, 2007.
16. Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, and Wei Hong. A Macroscope in the Redwoods. In *SenSys '05: Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 51–63, New York, NY, USA, 2005. ACM Press.
17. J. Tropp. *Topics in Sparse Approximation*. PhD thesis, University of Texas, Austin, 2004.