Scheduled Sampling for Robust Sensing

Noura Limam
D. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada

Malek Naouach
D. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada

***Abstract*—We consider the problem of optimizing the sensing strategy of a monitoring system in the presence of faulty sensors. We develop ORSg, an efficient data-driven algorithm for computing near-optimal sampling strategies to maximize the submodular utility of the sensing system with only a fraction of active and fault-prone sensors, whenever the utility function satisfies *submodularity*. Our approach combines techniques from information theory, game theory and submodular optimization. We empirically evaluate our algorithm on a real-world sensing problem.**

I. Introduction

There has been substantial work on energy conservation in the wireless sensors community [1], [2] and [3] and a number of energy saving models have been proposed to reduce power hungry operations. Adaptive sampling is a promising technique that exploits temporal and spatial correlation between measured samples to reduce the amount of data to be acquired. With adaptive sampling, samples are only acquired a fraction of the time at a fraction of the sensors. This has been shown to be particularly effective for classes of sensors, such as chemical and biological sensors, acoustic and seismic transducers, where the energy cost for data acquisition is significantly high.

Indeed acquiring data from a fraction of the sensors saves energy and prolongs the lifetime of the sensing system, however, it implies uncertainty about the values of the remaining ones. Besides, in addition to running out of power, active sensors may fail to communicate their data. Hardware failures, environmental hazards, extreme weather and ambient conditions negatively impact wireless communications and interfere with the quality of the collected data.

In this paper, we consider the problem of finding optimal sampling strategies in the presence of faulty sensors, and jointly address three design objectives; energy conservation, sensing utility maximization and robustness to sensor failures. We introduce an efficient data-driven algorithm for computing near-optimal sampling strategies that maximize the utility of the sensing system with only a fraction of active and fault-prone sensors, whenever the utility function satisfies *submodularity*. Our approach combines techniques from information theory, game theory and submodular optimization.

The remainder of this paper is organized as follows. Section II introduces the robust sampling problem and outlines our methodology. In Section III we describe the **ORS**g algorithm and prove its near-optimality. In Section IV we evaluate the algorithm empirically. Section V surveys relevant works. In Section VI we discuss possible extensions of the **ORS**g algorithm along with our future research directions.

II. The robust sampling problem

Let $V = \{s_1, ..., s_{|V|}\}$ be the set of sensors deployed in the sensing system. We denote by *sampling plan* any subset of sensors $A \in V$ that meets a given constraint such as on cardinality or cost. Let $\mathcal{P}(V)$ be the ground set of possible sampling plans.The robust sampling problem is about finding the sampling plan or a strategy over sampling plans in $\mathcal{P}(V)$ that best senses the condition assuming sensors are failure-prone.

We associate with each node $s \in V$ a discrete random variable $\mathcal{X}_s$. We want to select the most informative sampling plan $A \subset V$ to allow for an accurate estimation or prediction of the observed phenomenon. The most informative sampling plan should allow for little *uncertainty* about the remaining sensors $V - A$. This uncertainty can be captured by the *entropy* remaining in the random variable $\mathcal{X}_{V-A}$ conditioned by the readings of $\mathcal{X}_A$, known as the *conditional entropy* [4] $H(\mathcal{X}_{V-A}|\mathcal{X}_A)$. Hence our problem is to find the subset of sensors A such that:

$$A = \underset{A \in \mathcal{P}(V)}{\operatorname{argmax}} H(\mathcal{X}_{V-A}|\mathcal{X}_A)$$

In the following, $\bar{A}$ will denote $V - A$.

A. Problem Modelling

As sampling and communication failures may occur, only a fraction of the expected readings may be received and used to estimate the values of the other sensors. Because conditioning on less data increases the uncertainty in a random variable, robustness to sensor failures is a paramount criterion. The sampling strategy has to minimize this uncertainty in the worst case failure scenario.

The *robust sampling problem* can be seen as a *two-player zero-sum game*, i.e. a game in which one player's gain results only from the other player's equivalent loss. In this game, one player, called *MIN*, plays a sample plan A_i from $\mathcal{P}(V)$ while the opponent, called *MAX* plays a failure event from a set of possible events $\mathcal{F}$. Let m and n be the cardinality of $\mathcal{F}$ and $\mathcal{P}(V)$ respectively.

A fundamental theorem of game theory, the *Minimax theorem*, established by Neumann [5], states that every finite zero-sum two-person game, with a $n \times m$ game matrix, has (at

978-3-901882-66-1

least) a couple of optimal *mixed* strategies also called *Nash Equilibrium*; a probability assignment $x^* = (x_i^*)_{i=1..n}$ with $\sum_i x_i^* = 1$ over *MIN* 's possible strategies that will *minimize* her *maximum* loss, as well as a probability assignment $y^* = (y_j^*)_{j=1..m}$ with $\sum_j y_j^* = 1$ over *MAX* 's possible strategies that will *maximize* her *minimum* gain.

According to the Minimax theorem, the best outcome comes from assessing potential wins and losses, and developing a stochastic scheme for optimising the way that the players play. This suggests that the robustness and optimality of the sampling strategy is ensured provided that it is built on the probability distribution x^* over all possible plans A_i. Once x^* is computed, it is possible to set up a *schedule* among the sampling plans A_i, with each sample A_i operating at the frequency x_i^*. Indeed the sampling strategy could be established before hand which allows for the implementation of a *sample-sleep* schedule at the sensor level for better battery saving.

Now to find the optimal scheduling strategy we need to solve the game which requires solving a linear program (LP). Indeed it is possible to solve an LP in polynomial time, however this usually requires more than the number of variables of the LP raised to the third power operations [6]. Consider for instance that we want to sample from at most k sensors; i.e. any subset of sensors of cardinality less or equal to k is a potential sampling plan. The cardinality of $\mathcal{P}(V)$ is clearly exponential in the number of sensors. Thus, solving the LP is not tractable in general. In addition, even if the Nash strategy x^* can be computed "efficiently", it may be too complicated to implement. An optimal scheduling strategy is almost impractical if it has to randomize over a large number of simple sampling plans. This naturally leads us to the following question: instead of following a complex plan of actions over a very large number of simple sampling plans, can we approximate the optimal sampling strategy with a smaller number of plans?

B. Approximation and $\epsilon-$approximate Strategy with Small Support

The challenge here is to devise an efficient algorithm that computes a sampling strategy that achieves nearly as well as the optimal, and that randomizes over a *smaller* set of sampling plans. The problem is known as finding a near-optimal strategy with *small support*. Fortunately, results from [7] and [8] show that for any two-person zero-sum game there exists an $\epsilon-$equilibrium (assuming that the payoffs are in $[0, 1]$), i.e. a strategy for *MIN* for which the worst payoff is within an $\epsilon-$additive from optimum, with only a logarithmic support. Moreover the strategy of each player in such an equilibrium is uniform on a *small* multiset of pure strategies.

Theorem 1. *[7] Let X^l be all the mixed strategies that choose uniformly from a multiset of simple strategies $\mathcal{P}_l(V)$ of cardinality l. For any $\epsilon > 0$ and $l \geq [\frac{\ln m}{2\epsilon^2}]$, and considering the scaled payoff matrix Π;*

$$\max_{x \in X^l} \min_j \sum_i x_i \Pi(i,j) \leq \lambda^* + \epsilon$$

Equality holds only if $l = [\frac{\ln m}{2\epsilon^2}]$.

In a seminal result, Young [9] shows that a simple multiplicative update algorithm can be used to approximate optimal mixed strategies in an (arbitrary) matrix game, as long as 1) one of the players has a finite number of choices, and 2) the other player can compute best responses strategies -for the *MIN* player, the best response strategy to a given *MAX*'s strategy would be the strategy that minimizes MAX's gain-. In the following, we review this algorithm, adapted to the context of our application.

III. THE **ORS**G ALGORITHM

Without loss of generality, we consider in the following single sensor failure events; at each sampling round, at most one sensor will fail to sample or communicate with the rest of the system. The impact of communication failures can vary with the sensing application and particularly with the system's sensor-to-sink communication model. If the communications between the sensors and the sink are single-hop then the faulty sensor will only fail to deliver its readings to the sink. However if the communications between the sensors and the sink are multi-hop, the faulty sensor will fail to deliver all the readings it is forwarding to the sink. As a proof of concept we will assume a single-hop communication model. Indeed our methodology is generic enough to apply to any kind of failure events and any sensor-to-sink communication model.

Under the above conditions, it is easy to see that the set of possible failure events $\mathcal{F}$ is the same as V, thus $m = |V|$. The outcome of the pair of strategies (A_i, s_j) where the sampling plan A_i is selected to acquire data and the sensor $s_j \in A_i$ is selected to fail to aquire/deliver data, is the scaled entropy of $\bar{A}_i \cup \{s_j\}$ conditioned by the observations of $A_i - \{s_j\}$.

Using the identity $H(Y|X) = H(X,Y) - H(X)$, we have:

$$H(\mathcal{X}_{\bar{A}_i \cup \{s_j\}} | \mathcal{X}_{A_i - \{s_j\}}) = H(\mathcal{X}_V) - H(\mathcal{X}_{A_i - \{s_j\}})$$

Given V the (finite) set of sensors deployed to monitor a given condition, $\mathcal{P}(V)$ the set of possible sampling plans, the scaled uncertainty matrix $\big(\Pi(i,j)\big)$ where $\Pi(i,j)$ is the scaled expected uncertainty about the sensors in $\bar{A}_i \cup \{s_j\}$ conditioned by the observations in $A_i - \{s_j\}$, and $\epsilon > 0$, we consider the following Oblivious Randomized Sampling algorithm 1.

Algorithm 1 $ORS(V, \Pi)$

$\alpha \leftarrow e^{4\epsilon} - 1$
$\mathcal{P}^*(V) \leftarrow \{\}$ ▷ *Multiset of best response sensors sets*
$Y_j \leftarrow 1 (j = 1, ..., m)$
$X_i \leftarrow 0 (j = 1, ..., |\mathcal{P}(V)|)$
repeat
  choose $A_i \in \mathcal{P}(V)$ to minimize $\sum_j Y_j \Pi(i,j)$
  $X_i \leftarrow X_i + 1$
  $\mathcal{P}^*(V) \leftarrow \mathcal{P}^*(V) \cup \{A_i\}$ ▷ *A_i may appear more than once*
  $Y_j \leftarrow Y_j[1 + \alpha \Pi(i,j)] (j = 1, ..., m)$
until $|\mathcal{P}^*(V)| \geq \frac{\ln m}{2\epsilon^2}$ **return** $(x_i = \frac{X_i}{|\mathcal{P}^*(V)|})_{i=1,...,|\mathcal{P}^*(V)|}$

Theorem 2. *[9] The Oblivious Randomized Scheduling algorithm 1 returns an $\epsilon-$optimal solution to the robust sampling problem.*

At each iteration, the algorithm calls an oracle that returns the best response sensor set A_i to the probability distribution $(\frac{Y_j}{\sum_j Y_j})_{j=1..m}$ over failure events at s_j, and adds it to the multiset of best response sensors $\mathcal{P}^*(V)$.

The algorithm stops when the cardinality of the multiset reaches $\lceil \frac{\ln m}{2\epsilon^2} \rceil$. After $\lceil \frac{\ln m}{2\epsilon^2} \rceil$ iterations of the algorithm, the $\epsilon-$optimal solution is given by the returned vector $(x_i)_{i=1...|\mathcal{P}(V)|}$.

One important corollary that follows from the construction of the algorithm is that after $\lceil \frac{\ln m}{2\epsilon^2} \rceil$ iterations of the algorithm, the primal and the average dual values differ by at most ϵ. That is to say:

$$\max_j \sum_i x_i \Pi(i,j) \leq \sum_j y_j \sum_i x_i \Pi(i,j) + \epsilon \qquad (1)$$

A. *The* accelerated *Greedy Oracle*

The oracle in Algorithm 1 can be implemented as an exhaustive search over the set $\mathcal{P}(V)$ of all possible sensor sets. With real world problems and large scale sensing systems such a solution is expensive in general. In the following we show that a *Greedy*-based oracle not only has much lower complexity and execution time but also returns an approximate solution to $\underset{A_i \in \mathcal{P}(V)}{\operatorname{argmin}} \sum_j Y_j \Pi(i,j)$ with strong guarantees.

Let $h_j(A) = H(\mathcal{X}_{A-\{s_j\}})$ be the entropy of the readings received from the subset $A - \{s_j\}$ (assuming that s failed to deliver its reading).

Using the identity $H(\mathcal{X}_{V-A}|\mathcal{X}_A) = H(\mathcal{X}_V) - H(\mathcal{X}_A)$, we can see that minimizing the expected penalty $\sum_j Y_j \Pi(i,j)$ is the same as maximizing the expected entropy $\bar{h}(A_i) = \sum_j Y_j h_j(A_i)$, in other words;

$$\underset{A_i \in \mathcal{P}(V)}{\operatorname{argmin}} \sum_j Y_j \Pi(i,j) = \underset{A_i \in \mathcal{P}(V)}{\operatorname{argmax}} \sum_j Y_j h_j(A_i) \qquad (2)$$

Entropy is a positive and monotonically non decreasing function. The latter property follows from the *chain rule* $H(\mathcal{X}_A, \mathcal{X}_T) = H(\mathcal{X}_A|\mathcal{X}_T) + H(\mathcal{X}_T) = H(\mathcal{X}_T|\mathcal{X}_A) + H(\mathcal{X}_A)$, where the conditional entropy is itself positive.

Another major characteristic of the entropy is *submodularity* [10] [11]. We can easily show that for each j, $h_j(A) = H(X_{A-\{s_j\}})$ is also submodular, by realizing that $h_j(.)$ is the joint entropy function $H(.)$ defined on the ground set $\mathcal{P}(V - \{s_j\})$. The submodularity of $\bar{h}(A_i)$ follows as the class of submodular functions is closed under non-negative linear combinations.

Maximizing submodular functions is NP-hard in general. However, Nemhauser [12] proves that maximizing a monotone submodular function subject to a cardinality constraint admits a $(1 - 1/e)$ approximate solution, and assuming the $P \neq NP$ conjecture, no polynomial-time algorithm can achieve a better approximation ratio. Interestingly, the approximate solution is provided by the simple Greedy algorithm which starts with an empty solution $A = \varnothing$, grows A by successively adding some elements s from $V - A$ that maximize the marginal gain, and stops when size constraint is reached.

Considering a constraint k on the cardinality of the simple sampling plans A_i and by substituting to the exact oracle an oracle based on the simple Greedy selection rule (i.e. add the sensor s that maximizes the marginal gain $\sum_j Y_j[h_j(A \cup \{s\}) - h_j(A))]$), we are returned at each iteration of the algorithm a constant-factor approximate minimizer to $\sum_j Y_j \Pi(i,j)$ with substantially lower computational complexity, $\mathcal{O}(k|V|)$ in the number of sensors $|V|$ (if we consider the evaluations $\sum_j Y_j[h_j(A \cup \{s\}) - h_j(A))]$ for all $s \in V - A$ as atomic operation). To speedup the oracle, as suggested by Minoux [13], we can exploit the submodularity of the entropy function H a little further and reduce the number of evaluations $\sum_j Y_j[h_j(A \cup \{s\}) - h_j(A))]$ as follows:

Function 2 $GOracle(V, k, Y)$

```
A ← {}
for all s ∈ V do
    δ(s) ← ∞
repeat
    for all s ∈ V − A do
        eval_s ← FALSE
    break ← FALSE
    repeat
        s* ← argmax_{s∈V−A} δ(s)
        if eval_{s*} then
            A ← A ∪ {s*}
            break ← TRUE
        else
            δ(s*) = Σ_j Y_j[h_j(A ∪ {s*}) − h_j(A)]
            eval_{s*} ← TRUE
    until break
until |A| = k return A
```

We show in the following that the degree of approximation of the accelerated Greedy oracle (Algorithm 2) is carried over into the performance guarantees of the sampling algorithm.

B. ***ORS****g : The Oblivious Randomized Sampling Algorithm with Simple Greedy Oracle*

Algorithm 3 ORSg(V, k, ϵ)

```
α ← e^{4ε} − 1
P*_k(V) ← {}
Y_j ← 1   ∀j = 1, ..., m
repeat
    A* ← GOracle(V, H, k, Y)
    P*_k(V) ← P*_k(V) ∪ {A*}
    X_{A*} ← X_{A*} + 1                    ▷ # of occurrences of A*
    Y_j ← Y_j[1 + α(1 − h_j(A*)/H(X_V))](j = 1, ..., m)
until |P*_k(V)| ≥ ln m / 2ε² return (X_A / |P*_k(V)|)_{A∈P*_k(V)}
```

Theorem 3. *The* ***ORS****g algorithm is guaranteed to provide a sampling strategy that achieves at least a constant fraction of the optimal solution within $O(\frac{k|V| \ln m}{\epsilon^2})$ time.*

Proof. At any iteration the **ORS**g algorithm, the greedy oracle only returns an approximate optimizer A_{i^*}, such that

$\sum_j y_j h_j(A_{i^*}) \geq (1-\frac{1}{e})\max_{|A_i|\leq k}\sum_j y_j h_j(A_i)$, here $y_j = \frac{Y_j}{\sum_j Y_j}$. Given that $\sum_j y_j \Pi(i,j) = \frac{H(\mathcal{X}_V)-\sum_j y_j h_j(A_i)}{H(\mathcal{X}_V)}$, we have

$$\sum_j y_j \Pi(i^*,j) \leq (1-\frac{1}{e})\min_i \sum_j y_j \Pi(i,j) + \frac{1}{e} \quad (3)$$

Since at any step of the algorithm, the returned A_{i^*} is no longer an optimizer, it is easy to realize that the average of the values of the dual solution provided by the **ORS**g algorithm is such that $v_G \leq (1-\frac{1}{e})\sum_j y_j \min_i \Pi(i,j) + \frac{1}{e}$.

Accounting for Equation 1, it is easy to realize that **ORS**g returns a vector x such that:

$$\max_j \sum_i x_i \Pi(i,j) \leq (1-\frac{1}{e})\sum_j y_j \min_i \Pi(i,j) + \frac{1}{e} + \epsilon \quad (4)$$

Note that $\min_i \sum_j y_j \Pi(i,j)$ is a lower bound of $\lambda^* = \min_i \max_j \Pi(i,j)$. This implies that after $[\frac{\ln|V|}{2\epsilon^2}]$ iterations of **ORS**g, we have:

$$\max_j \sum_i x_i \Pi(i,j) \leq (1-\frac{1}{e})\lambda^* + \frac{1}{e} + \epsilon \quad (5)$$

For a given $0 < \rho < 1$ it suffices that $\epsilon \geq (1/e-\rho)\lambda^* - 1/e$ for the algorithm to compute a solution that is a $(1-\rho)$ fraction from optimal. The Theorem follows. □

IV. Experimental results

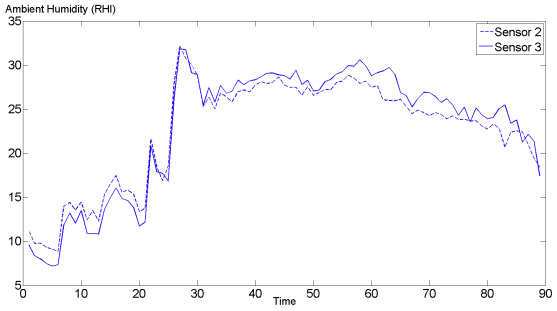
Our experimental evaluation is two-fold. First, we study the near-optimality of the **ORS**g algorithm by comparing the solution provided by **ORS**g to the optimal solution of the robust sampling problem obtained by solving the corresponding LP as discussed previously. Second, we consider an empirical evaluation of the sampling strategy selected by **ORS**g.

Our dataset consists of measurements of the ambient humidity collected every 15 minutes (in average) from 20 sensors distributed over the UC Berkeley Blue Oak Ranch Reserve. The dataset is available online through the eKoview web interface [1].

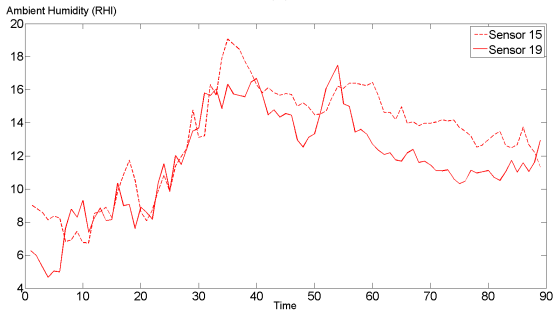
We consider the sensors records over a given day D. Figures 1 show the readings of sensors $s2$, $s3$, $s15$ and $s19$ between the time stamps $T = 0$ and $T = 90$. It is clear that records from $s2$ and $s3$ evolve similarly. In contrast, there is a discrepancy between the readings of $s15$ and $s19$. This suggests that $s2$ and $s3$ are more correlated than $s15$ and $s19$. Interestingly, we see that while humidity decreases in $s15$ (between $T = 5$ and $T = 8$), it increases in $s19$. This suggests that humidity is not isotonic across the field.

We build a Gaussian mixture model for the system fitted to day D's dataset with MATLAB's $gmdistribution.fit$ (Figure 2). We vary k, the constraint on the cardinality of the sampling plans and compare the near-optimal solution provided by **ORS**g to the optimal solution obtained by solving the LP. Interestingly, as shown in Figure 3, **ORS**g solves the problem very near-optimally; the solution provided by **ORS**g is much closer to the solution of the LP than to the guaranteed

[1] www.blueoakranchreserve.org



(a)



(b)

Fig. 1: (a) Highly correlated pair of sensors and (b) Weakly correlated pair of sensors

a priori bound. This suggests that a finer online bound should be investigated.

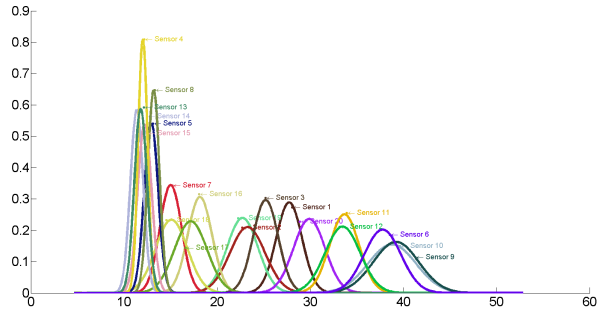


Fig. 2: Our model

Intuition suggests that due to changes in weather patterns, seasons, environment and other factors, the model is not time invariant. At some points in time, the system should collect readings from all the sensors to check the validity of the model and eventually build a new one. To determine a strategy for the adaptation procedure, key questions should be answered: Is there indeed a need for updating the sampling plan due to dynamic changes in the observed phenomenon? How often and under which condition the validity of the model should be checked? What should the conditions be that initiate the creation of the new model?

In order to test the validity of our sampling strategy we considered the following experiment. We consider the sam-

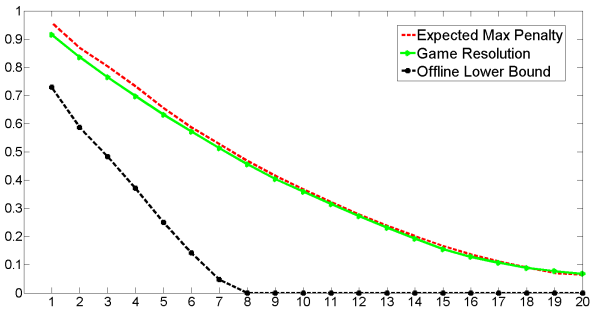


Fig. 3: Performance of $ORSg$ compared to the optimal solution

pling strategy Sch provided by **ORS**g for day D's model with $k = 4$ (i.e. only a quarter of the sensors are operated per sampling round). We calculate $H(X_s|X_{Sch})$ the expected uncertainties about each sensor in the system considering the measurements recorded by the sensors in Sch on day D, day $D+1$, day $D+1week$, and day $D+1month$. We calculate the actual entropy $H(X_s)$ of each sensor in the system on day D, day $D+1$, day $D+1week$, and day $D+1month$, and then the error $\frac{H(s)-H(X_s|X_{Sch})}{H(s)}$. Interestingly, as shown in Figures 4 the mean error does not exceed 0.45% over the 12 days that follow D. This suggests that the model and sampling strategy can be assumed valid for a fairly long time. Indeed it is possible to set a threshold beyond which the adaptation procedure should be triggered.

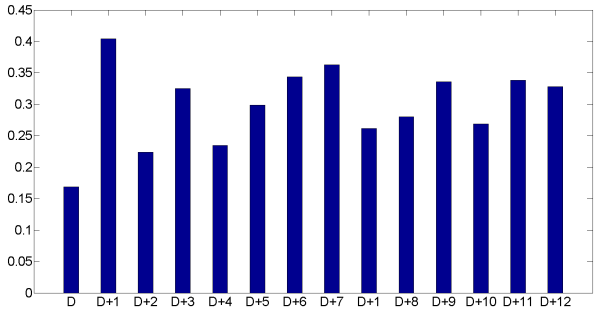


Fig. 4: Evaluation of the mean entropy error

V. Related Works

Early works on energy-effective sensing are on sensor deployment and coverage maximization [14] [15] [16] [17]. Coverage maximization has often been addressed in the literature with computational geometry-based approaches, which have been proven inefficient in many real-world scenarios [18]. In fact, geometry-based approaches typically assume that sensors have a fixed sensing radius. In practice, sensors make noisy measurements about the nearby environment, and their sensing area is not usually characterized by a regular disk. In contrast, our approach is data driven and does not place any assumption on the sensing coverage.

Our approach is stochastic and relies on statistical modeling. We exploit the characteristics of the sensed phenomenon in terms of random process and use a probabilistic model to predict the sensed value. Works like [19] [20] [21] [18] [22] well exemplify such approach.

In [19], an adaptive query-based sensing system is proposed. The sensed phenomenon is modeled as a time-varying multivariate Gaussian process. Historical data are used to build the initial probability density function over different attributes (temperature, humidity, voltage, etc.) at different sensors. The problem addressed in [19] is how to best sample from the sensors given correlations between the attributes at the different sensors.

In [20] [21] [18] sensor data is also assumed to follow a Gaussian distribution. Kernel linear regression is then used to predict the sensed condition in locations where no sensors are placed. The latter incorporates distance to predict the value at a given location. In [18] [21] the model is used to find a sensor placement that maximizes the mutual information between the selected placement and the uncovered locations. [21] optimizes the sensor placement while minimizing the communication cost in the sensor network. [22] considers sensor placement and scheduling jointly.

Even though the models used in [19] [20] [21] [18] [22] are very relevant to our work, the tackled problems are very different from our robust adaptive sampling problem. In fact, we further exploit the spatio-temporal correlations between measurements in order to reduce the number of data acquisitions and save energy. [23], and [24] are equally relevant from this perspective.

The main idea of [23] is that, nodes deployed with sufficient density do not have to sample the sensed field in a uniform way; more nodes have to be active in the regions where the variation of the measurements is high. Given a spatial distribution of the sensed phenomenon, the field is partitioned in a number of sub-squares with non-uniform resolution, where sensors are grouped in clusters. The sink will then activate additional sensors in the locations where the spatial correlation is low. To this end, it "backcasts" an activation message to the clusterheads who forward the received message to activate additional nodes in the cluster.

[24] addresses the energy conservation problem by adaptively coordinating the sleep schedules of sensors while guaranteeing that values of sleeping nodes can be recovered from the active nodes within a specified error bound. An isotonic regression model is used to characterize the sensed phenomenon and an adaptive sampling procedure is proposed. However, the proposed approach relies on the solution of a Mixed Integer Program to organize the sensors into domatic partitions such that each partition can predict the sensed condition on the entire network, which is intractable in general.

The approaches used in [23] and[24] are very different from ours, besides none of them considers robustness to failure events. The isotonicity constraint used in [24] and that

assumes that as the phenomenon being sensed increases, the sensors will experience an increase in their readings, does not hold in general. In fact we show in Figure 1 that as the humidity increases at one sensor, it decreases at the second. More generally, this assumption does not apply in large or heterogeneous settings.

VI. Discussion and Conclusion

We have devised a methodology for robust and energy efficient sensing by exploiting correlations in sensors data to reduce the number of data acquisitions by sampling at only a fraction of the sensors. We have presented and evaluated **ORS**g a low complexity, stochastic, data driven algorithm that solves the robust sampling problem to near-optimality with strong guarantees. The performance results are very promising. Possible extensions and future works include the design and implementation a *self-assessment/adaptation module* that assesses the validity of the schedule and the used model and triggers the adaptation of the system to the new environment.

References

[1] P. Berman, G. Calinescu, C. Shah, and A. Zelikovsky, "Efficient energy management in sensor networks," *Ad hoc and sensor networks*, vol. 2, pp. 71–90, 2005.

[2] L. Wang and Y. Xiao, "A survey of energy-efficient scheduling mechanisms in sensor networks," *Mobile Networks and Applications*, vol. 11, no. 5, pp. 723–740, 2006.

[3] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537–568, 2009.

[4] T. Cover and J. Thomas, *Elements of information theory*. Wiley-interscience, 2006.

[5] J. von Neumann, "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, vol. 100, no. 1, pp. 295–320, Dec. 1928. [Online]. Available: http://dx.doi.org/10.1007/bf01448847

[6] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, ser. STOC '84. New York, NY, USA: ACM, 1984, pp. 302–311. [Online]. Available: http://doi.acm.org/10.1145/800057.808695

[7] I. Althofer, "On sparse approximations to randomized strategies and convex combinations," *Linear Algebra and its Applications*, vol. 199, Supplement 1, no. 0, pp. 339 – 355, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0024379594903573

[8] R. J. Lipton, E. Markakis, and A. Mehta, "Playing large games using simple strategies," in *Proceedings of the 4th ACM conference on Electronic commerce*, ser. EC '03. New York, NY, USA: ACM, 2003, pp. 36–41. [Online]. Available: http://doi.acm.org/10.1145/779928.779933

[9] N. E. Young, "Randomized rounding without solving the linear program," in *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, ser. SODA '95. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1995, pp. 170–178. [Online]. Available: http://dl.acm.org/citation.cfm?id=313651.313689

[10] M. Madiman and P. Tetali, "Information inequalities for joint distributions, with interpretations and applications," *IEEE Trans. Inf. Theor.*, vol. 56, no. 6, pp. 2699–2713, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2046253

[11] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Information and Control*, vol. 39, no. 1, pp. 55 – 72, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S001999587891063X

[12] G. N. L. Wolsey; and M. Fisher, "An analysis of the approximations for maximizing submodular set functions," *Mathematical Programming*, 1978.

[13] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," in *Optimization Techniques*, ser. Lecture Notes in Control and Information Sciences, J. Stoer, Ed. Berlin/Heidelberg: Springer Berlin Heidelberg, 1978, vol. 7, ch. 27, pp. 234–243. [Online]. Available: http://dx.doi.org/10.1007/bfb0006528

[14] Z. Abrams, A. Goel, and S. Plotkin, "Set k-cover algorithms for energy efficient monitoring in wireless sensor networks," in *Proceedings of the 3rd international symposium on Information processing in sensor networks*, ser. IPSN '04. New York, NY, USA: ACM, 2004, pp. 424–432. [Online]. Available: http://doi.acm.org/10.1145/984622.984684

[15] C.-F. Huang and Y.-C. Tseng, "The coverage problem in a wireless sensor network," *Mob. Netw. Appl.*, vol. 10, no. 4, pp. 519–528, Aug. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1160162.1160175

[16] D. Tian and N. D. Georganas, "A coverage-preserving node scheduling scheme for large wireless sensor networks," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, ser. WSNA '02. New York, NY, USA: ACM, 2002, pp. 32–41. [Online]. Available: http://doi.acm.org/10.1145/570738.570744

[17] A. Chamam and S. Pierre, "On the planning of wireless sensor networks: Energy-efficient clustering under the joint routing and coverage constraint," *Mobile Computing, IEEE Transactions on*, vol. 8, no. 8, pp. 1077–1086, Aug.

[18] A. Krause, A. Singh, and C. Guestrin, "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies," *Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[19] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-based approximate querying in sensor networks," *The VLDB journal*, vol. 14, no. 4, pp. 417–443, 2005.

[20] *Distributed regression: an efficient framework for modeling sensor network data*, 2004. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1307317

[21] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Near-optimal sensor placements: Maximizing information while minimizing communication cost," in *Proceedings of the 5th International Conference on Information Processing in Sensor Networks (IPSN)*. New York, NY, USA: ACM Press, 2006, pp. 2–10. [Online]. Available: http://dx.doi.org/10.1145/1127777.1127782

[22] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin, "Simultaneous placement and scheduling of sensors," in *Proceedings of the 2009 International Conference on Information Processing in Sensor Networks*, ser. IPSN '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 181–192. [Online]. Available: http://dl.acm.org/citation.cfm?id=1602165.1602183

[23] R. Willett, A. Martin, and R. Nowak, "Backcasting: adaptive sampling for sensor networks," in *Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*. IEEE, 2004, pp. 124–133.

[24] F. Koushanfar, N. Taft, and M. Potkonjak, "Sleeping coordination for comprehensive sensing using isotonic regression and domatic partitions," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, april 2006, pp. 1 –13.

[25] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 420–429. [Online]. Available: http://doi.acm.org/10.1145/1281192.1281239