How Much Information Per Joule? Measuring the Energy Efficiency of Inferential Wireless Sensing

Paul G. Flikkema
EECS Department
Northern Arizona University
Flagstaff, AZ USA
Email: paul.flikkema@nau.edu

***Abstract*—Wireless sensor networks are increasingly seen not just as data collection instruments, but as part of an infrastructure designed to construct models of their embedding environments. Their energy supplies are limited, in turn constraining the amount of data that can be acquired and reported. This motivates consideration of the inferential value and energy cost of reporting the sensed data in the design of coding algorithms and/or reporting strategies, rather than the information content of the data itself. There is a need to evaluate the efficiency of these strategies and parametric variations of them in the context of model inference. This paper describes an information-theoretic measure, the relative divergence distribution (RDD), of their efficacy that relates inferential performance and the energy cost of reporting the data used in inference. It is non-asymptotic, and, as part of a Bayesian inference framework, does not require a prior distribution on the data model but can accommodate prior information on process model parameters. The inferential energy efficiency, based on the RDD, is a measure of the "fuel economy" of inferential sensing. These measures are applied to inferential sensing of a Bernoulli process, where it is demonstrated that the entropy of a data stream is not necessarily useful in understanding its value in inference.**

I. Introduction

In many application domains, wireless sensor networks are being called upon to gather data in a model-dependent context, where inference of a model for the embedding environment is at least as important as estimation of its current state. These process models and associated data models are often complex; fortunately, hierarchical Bayesian modeling and Markov Chain Monte Carlo (MCMC) techniques are powerful tools that together allow inference of complex process models from noisy spatio-temporal datasets, with quantification of state and parameter uncertainty via posterior distributions.

The paucity of available energy in most wireless sensing applications has motivated extensive effort in the wireless sensor networks research community, particularly on underlying principles and methods for compression of datasets and censoring of transmissions. However, the most energy efficient algorithms often degrade data fidelity and reduce model fidelity. These conflicting needs—energy conservation and fidelity—motivate exploration of the notion of inferential energy efficiency. How should the performance of a wireless sensor network be characterized? Traditional communication-theoretic measures, such as bit- and symbol-error probabilities, have the advantage of being independent of the type of data being transmitted. Their prevalence is rooted in the separation of source and channel coding (which is optimal for arbitrarily large block lengths), with source coding used to render each code symbol equally informative. When entropy rates are well-defined, network energy cost can be optimized using Slepian-Wolf coding for rate allocation [1]. However, wireless sensor network applications typically involve low data rates and low latency requirements so that block lengths are small. Moreover, they do not allow joint treatment of the value and energy cost of transmitted data in the context of inference.

What is needed is a useful quantitative measure of the energy efficiency of inferential sensing: how well a particular coding or reporting algorithm performs as a function of expended energy. The primary contribution of this paper is a means for measuring the performance of coding strategies in sensor networks in terms of the *relative divergence distribution* and a related summary measure, *inferential energy efficiency*, that precisely quantify the relationship between expended energy and the quality of the posterior.

II. Related Work and Overview

In many wireless sensing applications, transducer sampling intervals are selected based on a blend of intuition about the sampled processes and energy constraints (e.g., battery lifetime or power harvesting limitations). Source coding techniques and algorithms that censor or suppress reporting of data samples are then used to remove redundancy and save energy. They include distributed source coding [2], compressive sensing [3], quantization for estimation and classification [4], and joint coding and transmission control [5]. Deshpande et al. [6] suggested a query-oriented technique that used a one-step Markov multivariate Gaussian data model and Kalman filtering-based estimation. In [7], the sensor node infers an ARMA model that is used to respond to queries. Schemes for transmission censoring that use identical algorithms at the sensing node and the data center include [8], [9]; these also use a time series model. A technique that combines Bayesian

inference with randomized transmission control is described in [10]. The problem of discriminating between censored reports and reporting failures is treated in [11]. Statisticians and modelers have been developing powerful inferential approaches and applying them to state and parameter estimation in models that capture multiple sources of uncertainty [12], [13]. These approaches combine hierarchical Bayesian models with Markov chain Monte Carlo (MCMC) methods.

The key contributions of this paper are two measures of the effect of coding or reporting techniques on the energy efficiency of data/model inference. They are based on the Kullback-Leibler information divergence between two posterior distributions, which arises naturally in the setting that motivated the information bottleneck (IB). Developed in [14], the IB introduced the idea of the relevance of an encoded representation $\tilde{y}$ of a random quantity y in inferring a statistically related quantity of interest—here, the state x and parameters θ of a model. The goal is to construct a scheme so that $\tilde{y}$ preserves the maximum information about $\{x, \theta\}$. Applied to the problem of inference on wireless sensor networks, the IB-inspired approach developed here can be described as follows: Assume that a measurement y yields information about $\{x, \theta\}$ via the posterior $p(\theta, x|y)$. In a wireless sensor network, y is a vector-valued spatio-temporal dataset, specifically all the data samples acquired in a certain region over a period of time. We would like to know the effect of a strategy that yields a coded representation $\tilde{y}$ on (1) the quality of the posterior $p(\theta, x|\tilde{y})$ inferred from $\tilde{y}$ and (2) the energy cost of reporting that representation, *relative* to the best posterior $p(\theta, x|y)$ (i.e., when all data is reported). This relative quality is summarized using the Kullback-Leibler information divergence (KLD) of $p(\theta, x|\tilde{y})$ relative to $p(\theta, x|y)$. The KLD has been used to evaluate density estimation techniques for marginal posteriors generated using MCMC [15], and is a well-known conceptual tool for model selection and evaluation (see, e.g., [16]). Here, we use the KLD to compare posterior distributions and the corresponding energy cost of inferring them.

III. Preliminaries

Our sensing model is a simplified version of the model described in more detail in [17]; a short summary is presented here (see Figure 1). We assume that a sensor or group of sensors has taken a finite-length vector time series of measurements $y = (y_0, y_1, \cdots, y_{N-1})$ of the state x of a system parameterized by a vector θ. In practice, θ captures prior knowledge of the system via the prior distribution $p(\theta)$, and y may only represent a partial measurement of the state trajectory. The measurement y, which is assumed to be discrete-valued, is compressed to $\tilde{y}$ for transmission to a data center; this encoding $c(y) = \tilde{y}$ is modeled as the conditional distribution $p_c(\tilde{y}|y)$, which may be deterministic or random. Reporting may not be reliable ($z \neq \tilde{y}$); the effects of channel errors can be easily modeled via a mapping $p(z|\tilde{y})$, but perfect reporting is assumed in this paper.

In the Bayesian inference framework used here, the full posterior $p(y, x, \theta|\tilde{y})$ (for observations, state, and parameters) can be written as $p(y, x, \theta|\tilde{y}) = p(\tilde{y}, y, x, \theta)/p(\tilde{y})$ so that

$$p(y, x, \theta|\tilde{y}) \propto p(\tilde{y}, y, x, \theta). \tag{1}$$

Thus the posterior density is proportional to the joint distribution of $\{\tilde{y}, y, x, \theta\}$, which, noting that the ordered set $\{\theta, x, y, \tilde{y}\}$ is Markov, can be simplified to obtain

$$p(y, x, \theta|\tilde{y}) \propto p(\tilde{y}|y)p(y|x)p(x|\theta)p(\theta). \tag{2}$$

Even though practical models are often high-dimensional and hierarchical, Markov Chain Monte Carlo (MCMC) techniques for inference, including Gibbs and Metropolis-Hastings sampling, have proven to be highly effective in a broad array of applications.

Let the joint state and parameters be $\omega = \{x, \theta\}$. Note that the encoding $c(\cdot)$ and the observed dataset Y determine $\tilde{Y}$. With no coding, all data is sent and maximum energy is used; in this case, inference yields a posterior distribution that we denote as $\pi(\omega|Y)$, where

$$\pi(\omega|Y) \propto p(Y|x)p(x|\theta)p(\theta). \tag{3}$$

With encoding $c(\cdot)$, we obtain

$$p_c(\omega|Y) = \sum_{Y' \in \mathcal{S}_y} p(Y', \omega|\tilde{Y}), \tag{4}$$

where $\mathcal{S}_y$ is the sample space of y, and where $\tilde{Y}$ is implicit via the transformation $c(\cdot)$.

IV. A Performance Measure for Inferential Sensing

We compute the KLD from the best (full-data and thus maximum energy cost) posterior $\pi(\omega|Y)$ to the posterior $p_c(\omega|Y)$ inferred from the encoded data $\tilde{Y} = c(Y)$. We call this the *relative divergence*

$$D_c(Y) = D_{\mathrm{KL}}[\pi(\omega|Y) \parallel p_c(\omega|Y)] = E_\pi \log \frac{\pi(\omega|Y)}{p_c(\omega|Y)}; \tag{5}$$

the expectation is with respect to the posterior inferred from the full dataset; this is (at least in this setting) the reference distribution; however, the literature is not consistent on this labeling.

The expectation is in general an integral

$$D_c(Y) = \int \pi(\omega|Y) \log \frac{\pi(\omega|Y)}{p_c(\omega|Y)} d\omega \tag{6}$$

over the sample space of ω; this is unfortunately a high-dimensional space in most applications. However, in many cases a collection of marginal posteriors is desired and can be found using MCMC.

With base-2 logarithms, we can interpret (6) as the loss in fidelity from using the encoding $c(\cdot)$, measured as the number of additional bits that would be required, on average, to describe the reference posterior $\pi(\omega|Y)$ [18].

In our wireless sensing model, y is discrete, so that the average divergence is the weighted sum

$$D_c(y) = \sum_{Y \in \mathcal{S}_y} p(Y) D_c(Y), \tag{7}$$

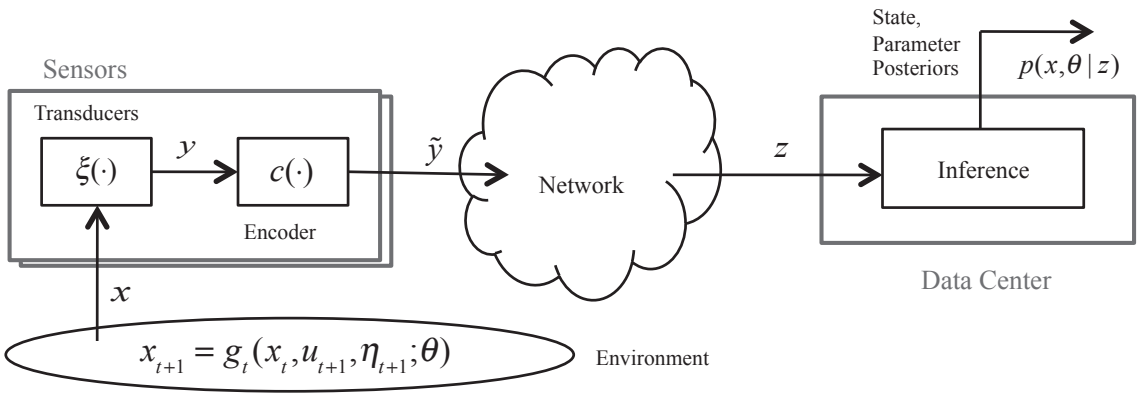

Fig. 1. Model of inferential wireless sensing. The process x_t capturing the state of the environment or its covariates are transduced, encoded, and sent to a data center. Model-based inference is performed on the received dataset.

which is also known as the conditional relative entropy [18] between the two posteriors.

When there is zero in-network energy expenditure, no data is reported and the only information available is the prior distribution $p(\theta)$. Here $\pi(\omega|Y)$ reduces to $\pi(\theta|Y)$, and the divergence is

$$D_0(Y) = D_{\mathrm{KL}}[\pi(\theta|Y) \parallel p(\theta)]. \tag{8}$$

Maximum energy is consumed when all data is sent (no encoding), resulting in the reference posterior $\pi(\omega|Y)$. Here the relative divergence is the singleton

$$D_A(Y) = D_{\mathrm{KL}}[\pi(\omega|Y) \parallel \pi(\omega|Y)] = 0 \; \forall \; Y \in \mathcal{S}_y. \tag{9}$$

When the prior distribution is non-informative and of finite support, it is uniform. Since the inference leans entirely on the data, the posterior is proportional to only the likelihood, so

$$D_c(Y) = D_{\mathrm{KL}}[\pi(\omega|Y) \parallel p(\tilde{Y}|\omega)/p(\tilde{Y})]. \tag{10}$$

showing that the divergence measure $D_c(\cdot)$ is implicitly a function of the prior.

Note that *any* finite-latency encoding/decoding scheme, including prediction- and transform-based approaches, can be evaluated using the relative divergence, since the divergence of the posterior distribution relative to the posterior for the uncoded case is computed. For example, the encoding may involve transformation to a domain (e.g., a feature space) that admits a sparse representation of the data. As described in [17], in some applications, interest is primarily in the strongly asymmetric case, where the encoding is extremely simple to minimize computational complexity and energy consumption at the encoding node. On the other hand, at the data center, decoding is absorbed into inference that exploits its comparatively unlimited computational power.

V. Inference and Energy

A. The Relative Divergence Distribution

For a given encoding algorithm $c(\cdot)$ and dataset realization Y, the reported information is $c(Y) = \tilde{Y}$ leading to the posterior $p_c(\omega|Y)$. The energy cost of reporting is also a function of $c(\cdot)$ and Y; we denote it as $\mathcal{E}_c(Y)$.

The Relative Divergence Distribution (RDD) $p[\mathcal{E}_c(Y), D_c(Y)]$ couples the relative divergence—the information gained due to reporting $c(Y)$—and $\mathcal{E}_c(Y)$. The RDD is a quantitative measure of the trade-off between energy savings and the quality of inference by capturing the effect of data collection on the posterior distribution and energy consumption.

The RDD is a joint probability distribution of the random variables $\mathcal{E}_c, D_c$. For example, when no data is reported, the RDD is over the set $\{(\mathcal{E} = 0, D_0(Y))\}_{Y \in \mathcal{S}_y}$. This is a set, not a point, since $\pi(\omega|Y)$ varies with Y in (8). More generally, consider the problem of reporting a correlated time series. The simplest approach might be to only report a fraction r of the data (using the native temporal basis), minimizing complexity at the sensor node. If r (and the associated subset of samples) is fixed, then r induces a specific posterior distribution and constant energy cost for all Y, so the RDD is a collection of distributions, each conditioned on a value of energy expenditure. More likely, the algorithm will be adaptive so that y induces a two-dimensional distribution in the energy-relative divergence plane. Section VI analyzes the RDD for what might be the canonical case where the entropy of the reported data $\tilde{y}$ and the associated energy cost are linear functions of r and hence directly related.

B. Energy Efficiency of Inference

Our concern is with the joint fidelity and the energy cost of encoding algorithms. Clearly, we can expend no energy on reporting by not sending data. The opposite extreme is to expend the maximum energy by reporting all the data; let $\mathcal{E}_A$ denote the energy cost in this case. Two questions arise: How do the relative divergence distributions of algorithms compare? And what is the relative performance of an algorithm across the range of energy expenditures from $\mathcal{E} = 0$ to $\mathcal{E}_A$? We introduce a measure based on the RDD called the inferential energy efficiency to quantitatively address these questions.

For a particular dataset, the total value of reporting the encoded data $c(Y) = \tilde{Y}$ is the resulting change in relative divergence from the zero energy case, i.e., $D_0(Y) - D_c(Y)$. The corresponding differential energy cost is $\mathcal{E}_c(Y)$ since $\mathcal{E}_0(Y) = 0$. We define the inferential energy efficiency (IEE) $\gamma_c(Y)$ for a dataset Y and an encoding scheme $c(\cdot)$ as the ratio of this change in divergence to the energy expenditure $\mathcal{E}_c(Y)$:

$$\gamma_c(Y) = \frac{D_0(Y) - D_c(Y)}{\mathcal{E}_c(Y)} \ \text{bits/J} \tag{11}$$

Note that γ_c is a random variable on the sample space of y, allowing rich characterization of its properties. A useful summary measure is the average inferential energy efficiency

$$E[\gamma_c] = \sum_Y p(Y)\gamma_c(Y) \ \text{bits/J}. \tag{12}$$

VI. Example: Inferential Sensing of a Bernoulli Process

To explore the RDD and inferential energy efficiency, we consider the problem of inferring θ for a Bernoulli(θ) process. This is a highly idealized model for inferential sensing of a binary-valued phenomena (such as presence/absence) over time or space. For example, a network might consist of widely-distributed nodes that periodically determine if the local ozone level exceeds a threshold, and θ would then be used in the construction of models for assessment of long-term public health effects. In this scenario, it is natural to consider the RDD as the number of observations increase; here, coding reduces to simply not reporting a fraction of a block of the observations to the data center, and hence the energy consumption is directly related to both the number of reported observations and the entropy of the reported data.

Let $|\cdot|$ denote the cardinality operator if the argument is a set, and the Hamming weight for a binary vector. Let the index set for any Y be $\mathcal{N} = \{1, \ldots, N\}$. For a block Y of N noiseless observations, the code $c(\cdot)$ selects a subset $\mathcal{C} \subseteq \mathcal{N}$ of indices that defines the vector $\tilde{Y} = c(Y) = (Y_i : i \in \mathcal{C})$ of $|\mathcal{C}| = \tilde{N} \leq N$ observations. In this case, the energy cost $\frac{\tilde{N}}{N}\mathcal{E}_A(Y)$ and the amount of information $\tilde{N}H_2(\theta)$ both scale with $\tilde{N}$, where $H_2(\cdot)$ is the binary entropy function.

Here we focus on parameter inference, so $\omega = \theta$. The model implies a very simple RDD: the energy increases linearly with $\tilde{N}$, and the divergence can be computed exactly from the appropriate distributions.

By independence, we have for any Y and $c(Y) = \tilde{Y}$ that

$$p_c(\theta|Y) = \frac{\theta^{|\tilde{Y}|}(1-\theta)^{\tilde{N}-|\tilde{Y}|}p(\theta)}{p(\tilde{Y})} \tag{13}$$

and

$$\pi(\theta|Y) = \frac{\theta^{|Y|}(1-\theta)^{N-|Y|}p(\theta)}{p(Y)}. \tag{14}$$

The structure of the posteriors admit fast computation; the number of computations required is $O(|\mathcal{C}|+1)(N-|\mathcal{C}|+1)+(N+1)$ (rather than $O(2^N)$) for each value of $\mathcal{E}_c$.

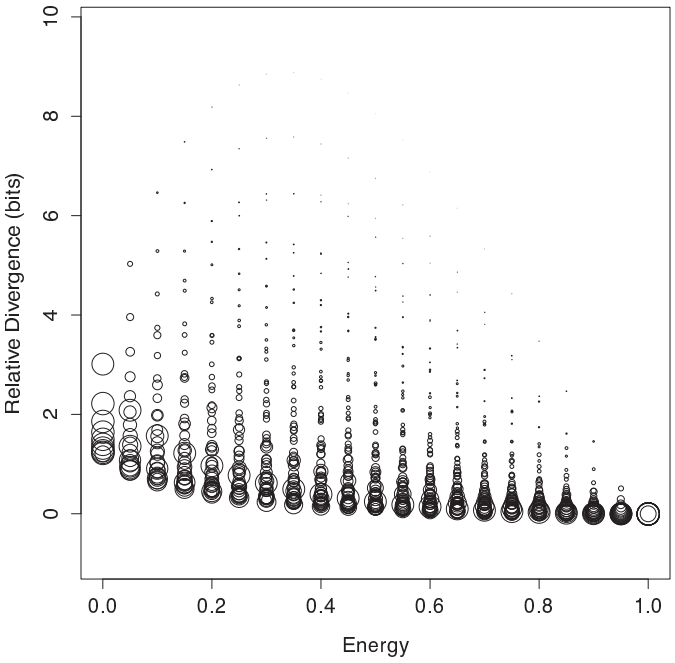

Fig. 2. Bubble plot of RDD for Bernoulli process with a maximum of $N = 20$ observations. Energy $\tilde{N}/N$ is relative to that used to transmit the full dataset Y. Bubble areas are scaled to probability at each energy level.

Figures 2 and 3 plot the RDD and the average inferential energy efficiency for the case of $N = 20$ observations and a prior $p(\theta)$ uniform on $(0, 1)$; energy is normalized ($\mathcal{E}_A(Y) = 1$). The RDD is a collection of probability mass functions (one for each discrete value of energy), as shown in Figure 2. The relative divergence for zero energy cost is simple because the KLD compares the uniform prior with the full-energy posterior. Even simpler is the point distribution $D_A(y)$ at $\mathcal{E}_A = 1$, as required by (9).

The average IEE (12) for this zero-latency case demonstrates the decreasing value/energy-cost ratio of reporting additional samples ($\circ$ symbols, Figure 3). The entropy of the reported data (like the energy cost) increases linearly with $\tilde{N}$, pointing out that inferential sensing requires different performance metrics than conventional data communication: in inferential sensing, the coding and reporting should distill the information in y relevant to the inference task.

The previous result applies to the case where each sample is reported as it is taken—the zero latency case—and reporting stops at sample $\tilde{N} \leq N$. The RDD and IEE also allow evaluation of the energy efficiency of coding under a fixed latency constraint. Assume that now a length-$\tilde{N}$ block $Y = (Y_1, \ldots, Y_{\tilde{N}})$ of a time series is collected (so that $\tilde{N}$ is the the reporting latency), following by coding. What is the RDD and IEE relative to the case when N samples are reported? For direct inference of θ, the sufficient statistic is the relevant information. When $\tilde{N}$ is known (e.g., via prior knowledge of the start of the block), the encoded data is $\tilde{Y} = \eta(Y) = |Y|$,

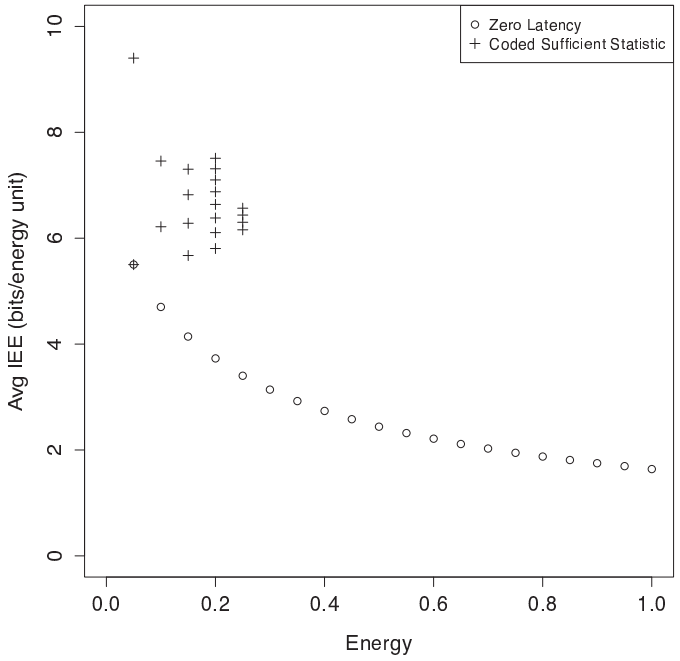

Fig. 3. Average inferential energy efficiency (IEE) $E[\gamma_c]$ (in bits per unit of energy) as a function of energy for reporting a Bernoulli process for a block of $N = 20$ outcomes. $\circ$ symbols are for zero-latency reporting with no coding; $+$ symbols are for encoding of the sufficient statistic under a latency constraint coinciding with the energy cost (see text). Energy is scaled as in Figure 2.

and the IEE is

$$\gamma_\eta(Y) = \frac{D_0(Y) - D_c(Y)}{\mathcal{E}_\eta(Y)} \tag{15}$$

(since $D_\eta(Y) = D_c(Y)$), and via straightforward fixed-length coding the corresponding normalized energy cost (relative to sending all N samples without coding) is, since $0 \leq |Y| \leq \tilde{N}$,

$$\mathcal{E}_\eta(Y) = \frac{\lceil \log \tilde{N} \rceil}{N}. \tag{16}$$

This is less than the energy $\tilde{N}/N$ required to report the block (observed state sequence). The result is that coding the sufficient statistic shifts the RDD leftward (not shown), and uses less energy, increasing energy efficiency ($+$ symbols in Figure 3). The cost is knowing only the sufficient statistic, rather than the state sequence, revealing again how inferential sensing can differ from conventional communication.

In general, the RDD gives a complete picture of the inferential loss of reporting a coded representation of the data, while the IEE measures the energy efficiency of the coding strategy. Both are functions of the energy required to send the encoded dataset.

VII. Conclusion

The relative divergence distribution (RDD) is a quantitative measure that couples the inferential performance of an encoding or reporting strategy to its energy use. It is a non-asymptotic Bayesian measure that clearly identifies the cost of reporting data, and can be used to evaluate any encoding or reporting scheme. The inferential energy efficiency summarizes the performance of encoding algorithms via a ratio of the improvement in relative divergence to the corresponding energy cost. Work is underway to apply these measures to encoding of real-world sensor network datasets for data and model inference.

Acknowledgment

The author thanks J. S. Clark and A. Gelfand for informative and stimulating discussions. This work was supported by NSF grant CNS-0540414.

References

[1] S. Baek, G. de Veciana, and X. Su, "Minimizing energy consumption in large-scale sensor networks through distributed data compression and hierarchical aggregation," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1130–1140, Aug. 2004.

[2] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[3] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. 5th Intl. Conf. on Information Processing in Sensor Networks (IPSN '06)*, 2006, pp. 134–142.

[4] R. M. Gray, "Quantization in task-driven sensing and distributed processing," in *Proc. IEEE ICASSP 2006*, 2006, pp. 1049–1052.

[5] S. L. Howard and P. G. Flikkema, "Progressive joint coding, estimation and transmission censoring in energy-centric wireless data gathering networks," in *5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS 2008)*, 2008, pp. 485–490.

[6] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *VLDB '04: Proceedings of the Thirtieth International Conference on Very Large Data Bases*, 2004, pp. 588–599.

[7] D. Tulone and S. Madden, "PAQ: time series forecasting for approximate query answering in sensor networks," in *3rd European Workshop on Wireless Sensor Networks (EWSN'06)*, 2006, pp. 21–37.

[8] C. Olston, J. Jiang, and J. Widom, "Adaptive filters for continuous queries over distributed data streams," in *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 2003, pp. 563–574.

[9] Y.-A. Le Borgne, S. Santini, and G. Bontempi, "Adaptive model selection for time series prediction in wireless sensor networks," *Signal Processing*, vol. 87, no. 12, pp. 3010–3020, 2007.

[10] G. Hartl and B. Li, "infer: A Bayesian inference approach towards energy efficient data collection in dense sensor networks," in *ICDCS '05: Proceedings of the 25th IEEE International Conference on Distributed Computing Systems*, Washington, DC, USA, 2005, pp. 371–380.

[11] A. Silberstein, G. Puggioni, A. Gelfand, K. Munagala, and J. Yang, "Suppression and failures in sensor networks: a Bayesian approach," in *VLDB '07: Proceedings of the 33rd International Conference on Very Large Data Bases*, 2007, pp. 842–853.

[12] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

[13] B. Carlin, J. Clark, and A. Gelfand, "Elements of hierarchical Bayesian inference," in *Hierarchical Modeling for the Environmental Sciences: Statistical Methods and Applications*, J. Clark and A. Gelfand, Eds. Oxford University Press, 2006.

[14] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Allerton Conference on Communication and Computation*, 1999, pp. 368–377.

[15] M.-H. Chen, Q.-M. Shao, and J. Ibrahim, *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag, 2000.

[16] K. Burnham and D. Anderson, "Kullback-Leibler information as a basis for strong inference in ecological studies," *Wildlife Research*, vol. 28, pp. 111–119, 2001.

[17] P. G. Flikkema, "Energy-efficient model inference in wireless sensing: Asymmetric data processing," in *IEEE Sensors 2010 Conference*, 2010.

[18] T. A. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.