

Scalable Algorithms for Distributed Statistical Inference

Animashree Anandkumar

Electrical & Computer Engr.,
Cornell University,
Ithaca, NY 14853, USA.
aa332@cornell.edu

ABSTRACT

The classical framework on distributed inference considers a set of nodes taking measurements and a fusion center making the final decision on the underlying phenomenon, without dealing with the issue of transporting the measurements to the fusion center. Such an approach introduces significant overhead in communication. Communicating all the raw data for inference is not scalable: in this case, the per-node average energy consumption and the total bandwidth requirement become unbounded as the network grows.

We design scalable algorithms for two scenarios with guarantees for inference whose communication requirements and complexity are bounded even as the network grows. This is achieved through distributed computation of a sufficient statistic, which results in reduction of data dimensionality while ensuring no loss in inference accuracy at the fusion center. The first scenario deals with multihop routing and fusion of spatially correlated measurements, incorporated through a Markov random field model. The second scenario deals with design of medium-access control (MAC) with the aim of computing a sufficient statistic for inference over a multiple access channel.

Categories and Subject Descriptors

H.1.1 [Systems and Information Theory]: Information theory

General Terms

Performance, Theory

Keywords

Statistical Inference, Scalable Algorithms, In-network Processing, Medium-Access Control.

This work was supported by the collaborative participation in Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011 and by the Army Research Office under Grant ARO-W911NF-06-1-0346 and by the IBM PhD fellowship 2008-09. The author is currently visiting Stochastic Systems Group at EECS Dept., MIT, Cambridge, MA, 02139.

1. INTRODUCTION

We are living in an increasingly networked world with networks of varying scales: the nodes in the network can comprise of billions of tiny devices, our personal mobile gadgets, or even our friends. The nature of links is also varied; they can be wireless, wire-line, or social links. There is rich interaction and information flow between these networks - for instance, between the computer and the social networks. So far, these different networks have been mostly studied as independent entities.

Another feature of these networks is the massive scale of the data they generate. Analysis of such large data sets requires scalable algorithms whose computational complexity does not grow with data. Moreover, since data is generated at a large number of nodes, the communication requirements of an algorithm is a key parameter. Depending on the application, algorithms need to undertake distributed computations at various nodes for communication requirements to be scalable in the data size and in the number of nodes in the network.

Many network applications involve collaborative processing of network data. For instance, in distributed statistical inference, the goal is to reach a decision about some common underlying phenomenon. Examples include intrusion detection, anomaly detection, temperature field estimation, and so on. We consider distributed inference where nodes communicate their data to a more powerful decision node called the fusion center, which then makes the final decision. We explicitly model the costs and constraints (e.g., energy, bandwidth) posed by the communication network to move data to the fusion center for inference.

If the nodes were to communicate all their raw data to the fusion centers, then such a scheme has a high communication cost, and is not scalable in the network size. Thus, it is important to “compress” the source data as much as possible. However, if the end goal is statistical inference, there is no need to communicate all the raw data. Instead, the collected network data often map into a *sufficient statistic* that consists of substantially fewer bits than the original data, while ensuring no loss in inference accuracy at the fusion center.

My thesis research considers schemes for reducing communication costs through distributed computation of the sufficient statistic over the communication network, based on the works in [1–9]. We consider two scenarios for the communication network. In the first scenario, we consider multi-hop routing with energy constraints, and develop in-network processing schemes for inference. In the second

scenario, we consider random access over a multiple access channel with energy and bandwidth constraints, and develop channel-aided computation schemes. These schemes are prime examples of cross-layer optimization in wireless networks. They couple the process of routing and medium access with the physical layer (where the energy expenditure occurs) and the application layer (where the statistical inference takes place). And they are examples of totally new and unexplored aspects of network operation. At the same time they raise some fundamental issues of communication costs for distributed computation and introduce trade-offs between communication and inference accuracy.

1.1 Multihop In-network Processing

Dependency graph is an effective model for describing relationships between nodes in a network based on some attribute, and needs to be inferred from the data generated by the nodes. For inference of the correct dependency graph model, the sufficient statistic has a compact form based on local dependency graph properties. In [1, 7, 9], we propose schemes for distributed computation of the sufficient statistic by exploiting the dependency graph structure.

Our scheme is scalable - it has strictly bounded average communication costs, even as the network grows, for a wide range of dependency graph models. Intuitively, when the dependency graph has only short-range edges between nearby nodes, the computation of the sufficient statistic can be undertaken locally with low communication costs. We provide a precise definition of such local dependency graphs based the concept of graph *stabilization* using the recent results on random geometric graphs. Such local dependency graphs occur in many scenarios - for example, the dependency between the location-based search queries and internet users; users near a particular location are more likely to query about that location than the ones further away. Another example is a sensor network measuring temperature of a field where nearby sensors tend to record similar temperatures.

We also provide a closed-form expression for average communication cost for inference under our scheme, and it has a nice representation in terms of the dependency graph, signal attenuation model and node placement. We use the expression to design efficient node placement strategies with low communication costs in [5]. We also address the related issue of selecting informative nodes for inference (sub-sampling) in [8] to further reduce the communication costs.

1.2 Medium-Access Control

We consider medium-access control (MAC) schemes for communication between the nodes in a network and the fusion center in [2–4]; the end goal is inference about a common underlying phenomenon measured by the nodes.

Traditionally, MAC schemes allocate transmission from different nodes to orthogonal channels (such as in time or frequency) to avoid interference. Instead, we propose a MAC scheme where nodes may interfere with one another, yet achieve good inference accuracy in the end. We allocate orthogonal channels to data levels: all nodes reaching the same local decision use the same orthogonal channel to transmit, if they decide to do so. This is an instance of channel-aided computation where we use the multiple access channel to compute a noisy histogram or the *type* of the local decisions, which serves as the sufficient statistic for inference. The bandwidth requirement of this scheme is independent

of the number of transmitting nodes, and is hence, scalable for large networks.

The extent to which interference aids inference depends on the nature of the multiple-access channel. Coherent channels add energy of the interfering signals more efficiently than canceling channels, and we quantify this behavior of the fading channels through a compact parameter, called the channel *coherence index*.

If the channel is canceling, then in our scheme, transmissions on independent orthogonal channels are more likely in order to avoid any interference. On the other hand, if the channel is coherent, simultaneous transmissions are more likely in our scheme. More specifically, we establish that for low coherence-index channels, our scheme has a finite optimal rate which maximizes inference performance. A sharp contrast is the extreme case when the channel is fully coherent (no random fading). In this case, we prove that the optimal rate is unbounded, which means that there should be simultaneous transmissions, in order to exploit the channel coherency.

Hence, our scheme adapts medium-access control based on the channel conditions to maximize inference performance, and it outperforms the classical orthogonal transmission scheme in terms of inference accuracy and bandwidth efficiency under the same energy budget.

Acknowledgement

The author thanks her advisor Prof. L. Tong, and her collaborators Prof. A.S. Willsky, Prof. J.E. Yukich, Dr. A. Swami and Prof. A. Ephremides, for their research inputs to the thesis.

2. REFERENCES

- [1] A. Anandkumar, A. Ephremides, A. Swami, and L. Tong. Routing for Statistical Inference in Sensor Networks. In S. Haykin and R. Liu, editors, *Handbook on Array Processing and Sensor Networks*, chapter 23. John Wiley & Sons, 2009.
- [2] A. Anandkumar and L. Tong. A Large Deviation Analysis of Detection over Multi-Access Channels with Random Number of Sensors. In *Proc. of ICASSP'06*, volume IV, pages 1097–1101, Toulouse, France, May 2006.
- [3] A. Anandkumar and L. Tong. Type-Based Random Access for Distributed Detection over Multiaccess Fading Channels. *IEEE Tran. Signal Proc.*, 55(10):5032–5043, Oct. 2007.
- [4] A. Anandkumar, L. Tong, and A. Swami. Distributed estimation via random access. *Information Theory, IEEE Transactions on*, 54(7):3175–3181, July 2008.
- [5] A. Anandkumar, L. Tong, and A. Swami. Optimal Node Density for Detection in Energy Constrained Random Networks. *IEEE Tran. Signal Proc.*, 56(10):5232–5245, Oct. 2008.
- [6] A. Anandkumar, L. Tong, and A. Swami. Detection of Gauss-Markov Random Fields with Nearest-neighbor Dependency. *IEEE Tran. Information Theory*, 55(2):816–827, Feb. 2009.
- [7] A. Anandkumar, L. Tong, A. Swami, and A. Ephremides. Minimum Cost Data Aggregation with Localized Processing for Statistical Inference. In *Proc. of INFOCOM*, pages 780–788, Phoenix, USA, April 2008.
- [8] A. Anandkumar, M. Wang, L. Tong, and A. Swami. Prize-Collecting Data Fusion for Cost-Performance Tradeoff in Distributed Inference. In *Proc. of IEEE INFOCOM*, Rio De Janeiro, Brazil, April 2009.
- [9] A. Anandkumar, J. Yukich, L. Tong, and A. Swami. Energy Scaling Laws for Distributed Inference in Random Networks. *Accepted to IEEE J. Selec. Area Comm., available on Arxiv*, Dec. 2008.