ERCIM Alain Bensoussan Fellowship Programme

ABCDE

MARIE CURIE ACTIONS

# Scientific Report

| | |
|---|---|
| First name / Family name of the *Fellow* | Mesrob I. Ohannessian |
| Nationality of the *Fellow* | Lebanese |
| Name of the *Host Organisation* | Inria Saclay Île-de-France |
| First name / Family name of the *Scientific Coordinator* | Pascal Massart |
| Period of the fellowship | 01/10/2012 to 30/09/2013 |

# I – SCIENTIFIC ACTIVITY DURING MY FELLOWSHIP

*Research Direction*

During my fellowship, I was interested in two extreme problems in statistics and machine learning: data scarcity and large datasets.

In situations with less data than traditionally assumed, one has to place structural assumptions in order to make statistical inference tasks well posed. In particular, I looked at rare probability estimation in settings where outcomes are discrete, as well as compression over large alphabets. In these problems, one natural structure is a tail characterization of the probability distribution. This notion is prevalent in continuous settings, namely tail probability estimation, but my work emphasizes that well-behaved tails are just as important in the discrete context.

As for large datasets, not only are efficient algorithms important to perform classical learning, but the additional data can be leveraged to improve performance. For example, selecting part of the data instead of the whole can help alleviate computational difficulty. Such pruning, however, induces artificial data scarcity, and it becomes necessary to understand the resulting trade-offs between data size and computational cost.

*Scientific Environment*

My fellowship was hosted within the SELECT project team of Inria. Their general vision of how models and structures have to be chosen, and the interaction between the algorithmic and statistical aspects of this inferential task were of great inspiration to me. One of the main technical tools used in model selection, the theory of concentration inequalities, was influential in various aspects of my thinking and research. This interest was maintained through periodic meetings with my scientific coordinator, Pascal Massart. At the Université Paris-Sud, where I was physically located, I worked closely with Elisabeth Gassiat. This was matched with equal interaction with Stéphane Boucheron, at Université Paris-Diderot. Also, as a result of the research exchange programme, I worked closely with Andreas Krause and his group, at ETH Zurich.

*Completed Projects*

The following are problems that I helped to concretely formulate and solve, and they led to completed manuscripts by the end of my fellowship (see next section). In particular, the first two were suggested in my research training programme.

- Universal compression with large alphabets, in particular leveraging an understanding of how a tail property governs the coding redundancy.
- Concentration properties of occupancy counts, in particular in wider settings than regular variation. Extensions to a functional version.
- Computation-statistics tradeoffs using data summarization, in particular characterizing regimes in which more data is or is not beneficial.

## II – PUBLICATIONS DURING MY FELLOWSHIP

*During my fellowship, the following papers were prepared and submitted:*

"**About Adaptive Coding on Countable Alphabets: Max-Stable Envelope Classes**"
With Stéphane Boucheron and Elisabeth Gassiat [Pending, Journal]

Universal source coding is, roughly, optimal compression without knowing the source but knowing a class it belongs to. An even more stringent property is adaptivity, performing well without even knowing the class, among a huge family of classes. Even mere universality is not possible without restrictions, and one typically assumes a finite alphabet with a known size. To model large alphabets, we instead propose classes with countably infinite alphabets characterized by a common dominating envelope. We give an explicit, computationally efficient, code that operates without knowing the envelope. It mixture-codes frequent symbols and integer-codes rare ones. Using regular variation theory and concentration of measure, we show that this code is near-adaptive to max-stable envelope classes. [Shortened abstract, invited to appear at ITA 2014 workshop.]

"**Computation-Statistics Tradeoffs using Coresets**"
With Mario Lučić, Amin Karbasi, and Andreas Krause [Pending, Conference]

Faced with massive data, is it possible to trade risk (statistical precision), space usage, and running time? This challenge lies at the heart of any large-scale learning problem. As a specific example, we consider k-means clustering. We show how we can strategically summarize the data (control space) in order to trade risk and time when data is generated by a probabilistic model. Our summarization is based on the notion of coresets from computational geometry. We show how summarization can be controlled such that for a fixed risk, computational time decays with the amount of data available. We also provide a simple meta-algorithm for navigating the space/time/risk tradeoff in practice. Our empirical results on real data sets demonstrate the existence and practical utility of such tradeoffs.

"**Estimating the Small Data in Big Data – Concentration and Regular Variation**"
With Anna Ben-Hamou and Stéphane Boucheron [Pending, Conference]

From the 20th century to the 21st, various disciplines have tried to infer something about scarcely observed events: zoologists about species, cryptologists about cyphers, linguists about vocabularies, and data scientists about almost everything. These problems are all about 'small data' within possibly much bigger data. Can we make such inference? As a concrete framework to studying such question we consider the occupancy problem in both its Bernoulli and Poisson settings. Our contribution is to give sharp concentration inequalities with explicit and reasonable constants, for the number of distinct observations, the missing mass, and the occupancy counts. We study the latter (in the Poisson setting) not only individually, but also as a vector over a (possibly growing) set of indices. We give both distribution-free and distribution-dependent results. When considering classes of distributions, we focus on those with a natural tail property given by regular variation.

*Additionally, the following continuation of my prior work was prepared and accepted to appear at the 2014 American Control Conference, June 4-6, 2014, in Portland, OR, USA:*

“**Dynamic Estimation of Price-Response of Deadline-Constrained Electric Loads with Threshold Policies**”
With Mardavij Roozbehani, Donatello Materassi, and Munther A. Dahleh [Accepted]

The paper presents a consistent and unbiased estimator for dynamic, one-step-ahead prediction of the aggregate response of a large number of individual loads to a common price signal, using only aggregate past response data. The price per unit of consumption is an exogenous signal which is updated at discrete time intervals. It is assumed that individual loads arrive in the system at random times with random demands and random consumption deadlines, and may defer their consumption up to the deadline in order to minimize their total cost. It is further assumed that the individual loads adopt a threshold policy in the sense that they only consume when the price is below a certain threshold. A dynamic aggregate model is constructed from models of independent individual loads. A consistent and unbiased estimator which only uses aggregate data, i.e., the price and aggregate consumption time-series is presented for estimating the aggregate consumption as a function of price.

## III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

*During the fellowship, I attended the following events, where I presented my work:*

- 43rd Saint-Flour Probability Summer School (43ème Ecole d’été de Probabilités), which took place July 7-20, 2013, in Saint-Flour, Cantal, France. The speakers were Krzysztof Burdzy "Brownian motion and its applications to mathematical analysis", Andrea Montanari "Statistical mechanics on random graphs", and Alexandre Tsybakov "Aggregation and high-dimensional statistics". The latter was particularly relevant to my work, through the concept of adaptivity.
- 15th IMS New Researchers Conference, which took place August 1-3, 2013, in Montréal, Québec, Canada. Following this event, I also attended parts of the 2013 Joint Statistical Meeting in Montréal, where I participated in the 2013 NISS/ASA Writing Workshop for Junior Researchers, on August 4, 2013.

*I attended the following workshops, without presentation on my part:*

- Fête Parisienne in Computation, Inference and Optimization: A Young Researchers' Forum. Held on March 20, 2013, at the Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette, France.
- Big data: Theoretical and Practical Challenges. Held on May 14-15, 2013, at the Institute Henri Poincaré, Paris, France.

*I also gave the following expository talks:*

- As part of the seminar series of the SELECT team of Inria, I gave a presentation on my doctoral work on November 8, 2012, at Université Paris-Sud, Orsay, France.
- As part of the seminar series of the Parisian machine learning community called SMILE, I gave a similar presentation on January 21, 2013, at the Ecole Normale Supérieure, Paris, France.

*Additionally, the following events were slated to happen after the end of my fellowship period, but with presentations that stem from my work during the fellowship:*

- Two talks on January 23, 2014, at Université Paris-Ouest, Nanterre, France: the first for a joint seminar of linguists and statisticians, "Mathématiques pour la linguistique de corpus", and the second, as part of the Modal'X seminar series.
- 2014 Information Theory and Applications Workshop, on February 9-14, San Diego, California, USA.

*Lastly, it is worth mentioning that I benefited from the various scientific meetings that happened regularly locally and in the Parisian region. These include: the SMILE machine learning seminars, the SemStats statistics seminars, the Point-de-Vue series at Paris-Diderot, the Laboratoire de Mathématiques d'Orsay colloquia and proba/stats seminars, and the SELECT project group meetings.*

## IV – RESEARCH EXCHANGE PROGRAMME

*I visited two other ERCIM institutes during my fellowship as part of the Research Exchange Programme (REP).*

- Andreas Krause at ETH Zurich in Zurich, Switzerland, on June 3-7, 2013.

  The purpose of the visit was to start collaboration on the topic of time-data tradeoffs in large data sets, especially in light of their expertise in a data summarization technique called corsets. During this visit, I had daily discussions with Andreas Krause and his group, and I also gave a presentation of some of my work during a special instance of their seminar series. After my visit, we continued our meetings via teleconferencing.

- Julien Hendrickx at Université Catholique de Louvain, Louvain-La-Neuve, Belgium, on September 11-13, 2013.

  The purpose of the visit was to start collaboration on the topic of networks of interacting agents, especially from the perspective of applications of rare probability estimation. I met Julien Hendrickx, his postdoc and other members of the ICTEAM institute. I also gave a presentation of my work during a joint group meeting. We plan to continue our collaboration in spring 2014.