Scientific Computation of Optimal Statistical Estimators

Houman Owhadi
CALIFORNIA INSTITUTE OF TECHNOLOGY 1200 E.
CALIFORNIA BLDV
PASADENA, CA 91125

07/13/2015
Final Report
# Scientific Computation of Optimal Statistical Estimators

## 1. Report Date (DD-MM-YYYY)
03-08-2015

## 2. Report Type
Final Technical

## 3. Dates Covered (From - To)
8/1/12 - 7/31/15

## 4. Title and Subtitle
Scientific Computation of Optimal Statistical Estimators

## 6. Author(s)
Houman Owhadi

## 5a. Contract Number
FA9550-12-1-0389

## 5b. Grant Number
FA9550-12-1-0389

## 5c. Program Element Number

## 5d. Project Number

## 5e. Task Number

## 5f. Work Unit Number

## 7. Performing Organization Name(s) and Address(es)
California Institute of Technology
1200 E. California Blvd.,
MC: 9-94
Pasadena, CA 91125

## 8. Performing Organization Report Number

## 9. Sponsoring/Monitoring Agency Name(s) and Address(es)
AF Office of Scientific Research
875 North Randolph Street, RM 3112
Arlington VA 22203

## 10. Sponsor/Monitor’s Acronym(s)
AFOSR

## 11. Sponsor/Monitor’s Report Number(s)

## 12. Distribution/Availability Statement
Distribution unlimited, approved for public release.

## 13. Supplementary Notes
Program Manager: Dr. Jean-Luc Cambier DR-04 USAF AFMC AFOSR/RTA2
Change in AFOSR program manager: Dr. Jean-Luc Cambier has replaced Dr. Fariba Fahroo AFOSR/RSL (now at DARPA)

## 14. Abstract
The past century has seen a steady increase in the need of estimating and predicting complex systems and making (possibly critical) decisions with limited information. Although computers have made possible the numerical evaluation of sophisticated statistical models, these models are still designed by humans because there is currently no known recipe or algorithm for dividing the design of a statistical model into a sequence of arithmetic operations. With the purpose of addressing this problem this program has developed (1) the foundations of a rigorous framework for the scientific computation of optimal statistical estimators/models and (2) the required calculus enabling the reduction of optimization problems over measures over spaces of measures and functions. Two highlights of the work accomplished consist of (1) the application of the calculus to the identification of brittleness in Bayesian inference and (2) the application of the framework to the automated identification of scalable linear solvers for PDEs with rough coefficients.

## 15. Subject Terms

## 16. Security Classification of:
- a. Report

## 17. Limitation of Abstract
- b. Abstract
- c. This Page

## 18. Number of Pages
7

## 19. Name of Responsible Person
Houman Owhadi

## 19b. Telephone Number (Include Area Code)
626-395-4547
INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.
Scientific Computation of Optimal Statistical Estimators

Final Report

Program Manager: Dr. Jean-Luc Cambier DR-04 USAF AFMC AFOSR/RTA2
Change in AFOSR program manager: Dr. Jean-Luc Cambier DR-04 USAF AFMC AFOSR/RTA2 has replaced Dr. Fariba Fahroo AFOSR/RSL (now at DARPA).
Contract/Grant #: FA9550-12-1-0389.
Reporting Period: 8/1/2012 to 7/31/2015.
PI: Houman Owhadi
Organization: California Institute of Technology

Abstract: The past century has seen a steady increase in the need of estimating and predicting complex systems and making (possibly critical) decisions with limited information. Although computers have made possible the numerical evaluation of sophisticated statistical models, these models are still designed by humans because there is currently no known recipe or algorithm for dividing the design of a statistical model into a sequence of arithmetic operations. With the purpose of addressing this problem this program has developed (1) the foundations of a rigorous framework for the scientific computation of optimal statistical estimators/models and (2) the required calculus enabling the reduction of optimization problems over measures over spaces of measures and functions. Two highlights of the work accomplished consist of (1) the application of the calculus to the identification of brittleness in Bayesian inference and (2) the application of the framework to the automated identification of scalable linear solvers for PDEs with rough coefficients.
Summary of the work accomplished

Enabling computers to think as humans have the ability to do when faced with uncertainty is challenging in several major ways: (1) Finding optimal statistical models remains to be formulated as a well posed problem when information on the system of interest is incomplete and comes in the form of a complex combination of sample data, partial knowledge of constitutive relations and a limited description of the distribution of input random variables. (2) The space of admissible scenarios along with the space of relevant information, assumptions, and/or beliefs, tend to be infinite dimensional, whereas calculus on a computer is necessarily discrete and finite. This program has laid down the foundations for addressing these challenges by developing the required framework and calculus.

A framework for the scientific computation of optimal models/estimators.
The framework, described in [14], consists of the full incorporation of computation and complexity into a natural generalization of Wald’s Statistical Decision Function framework [23, 24, 25, 26, 27] (based on a generalization of Von Neuman’s Theory of Games [21, 22]). In this framework optimal estimators/models are defined as optimal solutions of (minimax) adversarial games in which player A chooses the real system in an admissible set defined/constrained by available information and player B chooses the model/estimator, sees data generated by the real system and must predict some quantity of interest that is a function of the real system.

A calculus for manipulating infinite dimensional information structures.
The resolution of these minimax problems require, at an abstract level, searching in the space of all possible functions of the data. By restriction models to the Bayesian class, the complete class theorem [27, 1, 3, 19] allows to limit this search to prior distributions on the admissible set of candidates for the real system, i.e. to measures over spaces of measures and functions. To enable the computation of optimal estimators this program has therefore identified conditions under which minimax problems over measures over spaces of measures and functions can be reduced to the manipulation of finite-dimensional objects and developed the associated reduction calculus. For min or max problems over measures over spaces of measures (and possibly functions) this calculus can take the form of a reduction to a nesting of optimization problems over measures (and possibly functions for the inner part) [16, 11, 17], which, in turn, can be reduced to searches over extreme points [18, 20, 2, 13]. Specific applications and developments of this calculus are as follows. [2] has presented sufficient conditions under which an Optimal Uncertainty Quantification (OUQ, [18]) problem can be reformulated as a finite-dimensional convex optimization problem, for which efficient numerical solutions can be obtained. The sufficient conditions include that the objective function is piecewise concave and the constraints are piecewise convex. In particular, it has been shown that piecewise concave objective functions
may appear in applications where the objective is defined by the optimal value of a parameterized linear program. These developments have been applied in [2] to revenue maximization with stochastic supplies and to the optimal control of a power network with stochastic demands. The, more fundamental, results of [13, 12, 15] have laid down necessary steps for the identification of optimal reduced models on complex infinite-dimensional spaces (such reductions are ubiquitous with DFT and Navier-Stokes calculations). In particular, [13] has shown that, for the space of Borel probability measures on a Borel subset of a Polish metric space, the extreme points of the Prokhorov, Monge-Wasserstein and Kantorovich metric balls about a measure whose support has at most $n$ points, consist of measures whose supports have at most $n + 2$ points. Moreover, using the Strassen and Kantorovich-Rubinstein duality theorems [13] has developed efficiently computable supersets of the extreme points. [12] has shown that for a Gaussian measure on a separable Hilbert space, the family of conditional measures associated with conditioning on a closed subspace are Gaussian with covariance operator the short of the covariance operator to the closed subspace. [15] has demonstrated that a reproducing kernel Hilbert space of functions on a separable absolute Borel space or an analytic subset of a Polish space is separable if it possesses a Borel measurable feature map.

**Bayesian Brittleness.** In the process of its development this calculus has been applied to analyse the robustness of Bayesian Inference under finite information [16, 11, 17, 10]. This analysis has uncovered the possible extreme sensitivity (brittleness) of Bayesian inference (in the TV and Prokhorov metrics or for Bayesian models that exactly capture an arbitrarily large number of finite-dimensional marginals of the data-generating distribution) and suggested that robust inference, in a continuous world under finite-information, should be done with reduced/coarse models rather than highly sophisticated/complex models (with a level of coarseness/reduction depending on the available finite-information) [17]. More precisely, although Bayesian methods are robust when the number of possible outcomes is finite or when only a finite number of marginals of the data-generating distribution are unknown, they appear to be generically brittle when applied to continuous systems (and their discretizations) with finite information on the data-generating distribution. Furthermore, if closeness is defined in terms of the total variation metric or the matching of a finite system of generalized moments, then (1) two practitioners who use arbitrarily close models and observe the same (possibly arbitrarily large amount of) data may reach opposite conclusions; and (2) any given prior and model can be slightly perturbed to achieve any desired posterior conclusions. The mechanism causing brittleness/robustness suggests that learning and robustness are antagonistic requirements and raises the question of a missing stability condition for using Bayesian Inference in a continuous world under finite information.
**Automation of the process of scientific discovery.** In the process of developing this framework and calculus this program has started addressing (as a direct application of the framework and calculus) the fundamental question of whether scientific discovery can be computed, i.e., can the process of scientific discovery by guided by, or turned into an algorithm? (in some sense this question is related to that of whether machines can think). This program has addressed three notoriously difficult examples in which the answer to the above question is positive. The first one concerns the identification of new Selberg integral formulae [11] (a notoriously difficult problem of pure mathematics that has been turned into an algorithm). The second one concerns the identification of accurate, localized bases for numerical homogenization/coarse graining with optimal recovery properties [8] (a notoriously difficult problem of applied mathematics that has been turned into an algorithm). And the third one concerns the identification of near-linear complexity linear numerical solvers [9] (a notoriously difficult CSE problem that has been turned into an algorithm).

**Gamblets.** This latter example has lead the to the discovery of Gamblets [9] and shown that the discovery/design of scalable numerical solvers can be addressed/automated as a UQ problem by reformulating the process of computing with partial information and limited resources as that of playing underlying hierarchies of adversarial information games. As an illustration [9] has shown how the application of the proposed approach to the resolution of elliptic PDEs with rough coefficients leads to a near-linear complexity multigrid/multiresolution method with rigorous a-priori accuracy and performance estimates. In this application, the numerical solver has been discovered by identifying optimal strategies for gambling on the value of the solution of the PDE based on hierarchies of nested measurements of its solution or source term.

**Development an efficient framework for heterogeneous computing and robust optimization** This program has continued the development of (1) a computational job management framework (*pathos*) (a parallel graph execution framework providing a high-level programmatic interface to high-performance computing [http://trac.mystic.cacr.caltech.edu/project/pathos, [5]]) that offers a simple, efficient, and consistent user experience in a variety of heterogeneous environments from multi-core workstations to networks of large-scale computer clusters and (2) a robust optimization framework (*mystic*) (a highly-configurable optimization framework [http://trac.mystic.cacr.caltech.edu/project/mystic, able to drive material science code to fit structures [6, 4]]) that incorporates the mathematical framework described in [18, 7], and has provided an interface to prediction, certification, and validation as a framework service.

More precisely, under this program, asynchronous computing capabilities were added to pathos. Worker pools now provide asynchronous maps and pipes, as well as iterative ordered and unordered asynchronous variants. New asynchronous conditional parallel maps were added, which are both robust against failure and potentially
orders of magnitude faster than blocking maps. Conditional maps terminate when
the desired (potentially statistical) condition is met, as opposed to waiting for all
results to return. The klepto package was created to provide an abstraction for stor-
age and retrieval of objects in a database, in memory, or on disk. klepto provides
asynchronous and distributed parallel caching (as opposed to recalculation) and cache
interpolation strategies, and can be used to decouple the workflow and management
of ASGs that span distributed resources. klepto also provides hierarchical caching, so
for example a fast local cache could be maintained in memory with the most recently
used entries, while a centralized global database serves as a second tier for all entries
not interpolated or found in the fast local cache.

The majority of mystic was converted to asynchronous computing, thus enabling
optimization to dramatically scale in size and complexity. Optimizers in mystic can
now proceed in a step-by-step iterative fashion, potentially saving state at each step.
This change enables mystic’s optimizers to serve as a long-running daemon process
that dynamically responds to new information – essentially optimizers have been
converted to provide a ”streaming” or ”event” mode, to tackle real-time updates of
information about the constraints or the cost function.

Given enough parallel resources, mystic’s ensemble solvers demonstrate orders
of magnitude improvements in speed and accuracy over industry standard genetic
algorithms. With the addition of klepto, mystic’s ensemble solvers were augmented
to provide N-dimensional global search capabilities. For example, parallel ensembles
of optimizers can be launched to search for all critical points and inflection points of
an unknown surface, terminating only after no further points are found. The resulting
points can then be fed into an N-dimensional interpolation engine, to produce a fast
accurate surrogate model for the unknown surface.

Broader impact of the work accomplished. H. Owhadi and C. Scovel have
been interviewed by HPC Wire\footnote{See \url{www.hpcwire.com/2013/09/13/the_masters_of_uncertainty/}}. The Bayesian Brittleness papers have generated
significant blog activity\footnote{See for instance \url{http://errorstatistics.com/2015/01/08/on-the-brittleness-of-bayesian-inference-an-update-owhadi-and-scovel-guest-post/}}. Gamblets have been presented at a plenary lecture at SIAM
CSE 2015\footnote{See \url{https://www.pathlms.com/siam/courses/1043/sections/1259/thumbnail_video_presentations/9883}}. H. Owhadi is co-editing “the Handbook of Uncertainty Quantification”
(Springer) with R. Ghanem and D. Higdon. M. McKerns is editing a chapter in that
book (on software aspects). H. Owhadi has been invited (by Dr. Bruce Suter DR-04
USAF AFMC AFRL/RITB) to AFRL, Rome NY to present and discuss the results of
\cite{Owhadi2013}. Schlumberger is exploring the incorporation of the results of \cite{Owhadi2013} into its subsurface
flows software. Gamblets have lead to a provisional patent (number 62/130,374).
References


DISTRIBUTION A: Distribution approved for public release.


1. Report Type
Final Report

Primary Contact E-mail
Contact email if there is a problem with the report.
owhadi@caltech.edu

Primary Contact Phone Number
Contact phone number if there is a problem with the report
8182757899

Organization / Institution name
Caltech

Grant/Contract Title
The full title of the funded effort.
Scientific Computation of Optimal Statistical Estimators

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".
FA9550-12-1-0389

Principal Investigator Name
The full name of the principal investigator on the grant or contract.
Houman Owhadi

Program Manager
The AFOSR Program Manager currently assigned to the award
Dr. Jean-Luc Cambier

Reporting Period Start Date
08/01/2012

Reporting Period End Date
07/31/2015

Abstract
The past century has seen a steady increase in the need of estimating and predicting complex systems and making (possibly critical) decisions with limited information. Although computers have made possible the numerical evaluation of sophisticated statistical models, these models are still designed \emph{by humans} because there is currently no known recipe or algorithm for dividing the design of a statistical model into a sequence of arithmetic operations. With the purpose of addressing this problem this program has developed (1) the foundations of a rigorous framework for the scientific computation of optimal statistical estimators/models and (2) the required calculus enabling the reduction of optimization problems over measures over spaces of measures and functions. Two highlights of the work accomplished consist of (1) the application of the calculus to the identification of brittleness in Bayesian inference and (2) the application of the framework to the automated identification of scalable linear solvers for PDEs with rough coefficients.

Distribution Statement
This is block 12 on the SF298 form.
Distribution A - Approved for Public Release

Explanation for Distribution Statement
DISTRIBUTION A: Distribution approved for public release.
If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form
Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF. The maximum file size for an SF298 is 50MB.

SF 298 Form Final Technical.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF. The maximum file size for the Report Document is 50MB.

finalreport.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Changes in research objectives (if any):
None

Change in AFOSR Program Manager, if any:
Dr. Jean-Luc Cambier DR-04 USAF AFMC AFOSR/RTA2 has replaced Dr. Fariba Fahroo AFOSR/RSL (now at DARPA)

Extensions granted or milestones slipped, if any:
None

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

<table>
<thead>
<tr>
<th>Category</th>
<th>Starting FY</th>
<th>FY+1</th>
<th>FY+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equipment/Facilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Report Document

Report Document - Text Analysis

DISTRIBUTION A: Distribution approved for public release.
2. Thank You

E-mail user

Aug 13, 2015 13:20:05 Success: Email Sent to: owhadi@caltech.edu