Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 326 (2017) 622–641
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Data Driven Computing with noisy material data sets

T. Kirchdoerfer, M. Ortiz*

Graduate Aerospace Laboratories, California Institute of Technology, 1200 E. California Blvd., MC 105-50, Pasadena, CA, 91125, USA

Received 3 March 2017; received in revised form 17 May 2017; accepted 27 July 2017
Available online 24 August 2017

Abstract

We formulate a Data Driven Computing paradigm, termed max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing and is robust with respect to outliers. Robustness is achieved by means of clustering analysis. Specifically, we assign data points a variable relevance depending on distance to the solution and on maximum-entropy estimation. The resulting scheme consists of the minimization of a suitably-defined free energy over phase space subject to compatibility and equilibrium constraints. Distance-minimizing Data Driven schemes are recovered in the limit of zero temperature. We present selected numerical tests that establish the convergence properties of the max-ent Data Driven solvers and solutions.

Keywords: Data science; Big data; Approximation theory; Scientific computing

1. Introduction

Despite the phenomenal growth of scientific computing over the past 50 years, several stubborn challenges have remained foci of extensive research to this day. One of those challenges is *material modeling*. The prevailing and classical scientific computing paradigm has been to calibrate empirical material models using observational data and then use the calibrated material models in calculations. This process of modeling inevitably adds error and uncertainty to the solutions, especially in systems with high-dimensional phase spaces and complex material behavior. This modeling error and uncertainty arises mainly from imperfect knowledge of the functional form of the material laws, the phase space in which they are defined, and from scatter and noise in the experimental data. Simultaneously, advances in experimental science over the past few decades have changed radically the nature of science and engineering from *data-starved* fields to, increasingly, *data-rich* fields, thus opening the way for the application of the emerging field of *Data Science* to science and engineering. Data Science currently influences primarily non-STEM fields such as marketing, advertising, finance, social sciences, security, policy, and medical informatics, among others. By contrast, the full potential of Data Science as it relates to science and engineering has yet to be explored and realized.

* Corresponding author.
E-mail address: ortiz@caltech.edu (M. Ortiz).

http://dx.doi.org/10.1016/j.cma.2017.07.039
0045-7825/

The present work is concerned with the development of a Data Science paradigm, to be referred to as *Data Driven Computing*, tailored to scientific computing and analysis, cf. [1]. Data Driven Computing aims to formulate initial–boundary-value problems, and corresponding calculations thereof, directly from material data, thus bypassing the empirical material modeling step of traditional science and engineering altogether. In this manner, material modeling empiricism, error and uncertainty are eliminated entirely and no loss of experimental information is incurred. Here, we extend earlier work on Data Driven Computing [1] to random material data sets with finite probability of *outliers*. We recall that the Data Driven Computing paradigm formulated in [1], or *distance-minimizing* Data Driven Computing, consists of identifying as the best possible solution the point in the material data set that is closest to satisfying the field equations of the problem. Equivalently, the distance-minimizing Data Driven solution can be identified with the point in phase space that satisfies the field equations and is closest to the material data set. It can be shown [1] that distance-minimizing Data Driven solutions converge with respect to uniform convergence of the material set. However, distance-minimizing Data Driven solutions can be dominated by *outliers* in cases in which the material data set does not converge uniformly. Distance-minimizing Data Driven solvers are sensitive to outliers because they accord overwhelming influence to the point in the material data set that is closest to satisfying the field equations, regardless of any clustering of the material data points.

The central objective of the present work is to develop a new Data Driven Computing paradigm, to be called max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing and is robust with respect to outliers. Robustness is achieved by means of clustering analysis. Specifically, we assign data points a variable relevance depending on distance to the solution and through maximum-entropy estimation. The resulting scheme consists of the minimization of a free energy over phase space subject to compatibility and equilibrium constraints. We note that this problem is of non-standard type, in that the relevant free energy is a function of state defined over phase space, i.e., a joint function of the driving forces and fluxes of the system. Max-ent Data Driven solutions are robust with respect to outliers because a cluster of data points can override an outlying data point even if the latter is closer to the constraint set that any point in the cluster. The distance-minimizing Data Driven schemes [1] are recovered in the limit of zero temperature. We also develop a simulated annealing scheme that, through an appropriate annealing schedule zeros in on the most relevant data cluster and the attendant solution. We assess the convergence properties of max-ent Data Driven solutions and simulated annealing solver by means of numerical testing.

The paper is organized as follows. In Section 2, we begin by laying out the connection between Data Science and Scientific Computing that provides the conceptual basis for Data Driven Computing. In Section 3, we turn attention to random material data sets that may contain outliers, or points far removed from the general clustering of the material data points, with finite probability and develop max-ent Data Driven solvers by an appeal to Information Theory and maximum-entropy estimation. In Section 4, we develop a simulated annealing solver that zeros in on the solution, which minimizes a suitably-defined free energy over phase space by progressive quenching. In Section 5, we present numerical tests that assess the convergence properties of max-ent Data Driven solutions with respect to uniform convergence of the material data set. We also demonstrate the performance of Data Driven Computing when the material behavior itself is random, i.e., defined by a probability density over phase space. Finally, concluding remarks and opportunities for further development of the Data Driven paradigm are presented in Section 6.

2. The data driven science paradigm

In order to understand the hooks by which Data Science may attach itself to Scientific Computing, it helps to review the structure of a typical scientific calculation. Of special import to the present discussion is the fundamentally different roles that conservation and material laws play in defining that structure, with the former setting forth hard universal or material-independent constraints on the states attainable by the system and the latter bringing in material specificity open to empirical determination and sampling.

2.1. The 'anatomy' of boundary-value problems

We begin by noting that the field theories that provide the basis for scientific computing have a common general structure. Perhaps the simplest field theory is potential theory, which arises in the context of Newtonian mechanics, hydrodynamics, electrostatics, diffusion, and other fields of application. In this case, the field φ that describes the global state of the system is scalar. The *localization law* that extracts from φ the *local state* at a given material point is $\epsilon = \nabla\varphi$, i.e., the localization operator is simply the gradient of the field, together with essential boundary conditions