Data Driven Computing

Thesis by

Trenton Kirchdoerfer

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2018

(Defended July 12, 2017)

Trenton Kirchdoerfer

ORCID: [0000-0003-2290-1857]

For Yuka: Herein lies the foundations of our Mountain-Village

Acknowledgments

First thanks go to my adviser Michael Ortiz, who provided guidance, showed patience and shared experience all to my benefit. His demonstrated research strategies have embodied a fundamental way of examining problems I will forever associate with him, Caltech and the advancement of science. I am extremely grateful to have been able to work with so able a mentor. I am also thankful to the many members of the Ortiz group, including Lydia and Marta, with whom I have frequently discussed life and work. In this vein special thanks are owed to Amuthan, who opened his door to my questions and requests for aide too many times to count. I remain indebted to the many other friends and classmates who were pivital in helping me navigate the difficulties of graduate life. I also recieved a great deal of guidance and advise from my coworkers and managers at the Southwest Research Institute, which provided a foundation that was an important part of my success here at Caltech.

The family support I recieved these four years has been amazing. Both my wife's and my own parents were supportive from the start, including emotional, financial, and travel support. They remained voices of optimism and encouragement throughout the years this degree required, never once questioning our decision to return to school from a paying career. To my children, Takato and Masato, you were shining lights at the end of many dark days; many times your smiles and simple joys provided me perspectives that would ease whatever deep concerns I was having. Finally I must thank my wife Yuka for supporting me so completely during my studies here at Caltech. Yuka, I took classes and explored research, while you shouldered the heavy burdens of giving birth and raising children. None of the work I have done would have been possbile without your devotion and dedication to our family.

Abstract

Data Driven Computing is a new field of computational analysis which uses provided data to directly produce predictive outcomes. This thesis first establishes definitions of Data-Driven solvers and working examples of static mechanics problems to demonstrate efficacy. Significant extensions are then explored to both accommodate noisy data sets and apply the deveoloped methods to dynamic problems within mechanics. Possible method improvements discuss incorporation of data quality metrics and adaptive data sampling, while new applications focus on multi-scale analysis and the need for public databases to support constitutive data collaboration.

Published Content and Contributions

1. Kirchdoerfer, T., and Ortiz, M. Data-driven computational mechanics. *Computer Methods in Applied Mechanics and Engineering 304* (2016), 81–101,DOI:10.1016/j.cma.2016.02.001
   T.K. Implemented all code and method demonstrations, also suggested the topology of convergence for R^3 continuum mechanics extension.

2. Kirchdoerfer, T., and Ortiz, M. Data driven computing with noisy material data sets. *Computer Methods in Applied Mechanics and Engineering* (Submitted March 2017)
   T.K. Implemented all code and method demonstrations presented and added annealing control parameter λ adjust annealing schedule.

3. Kirchdoerfer, T., and Ortiz, M. *Data-Driven Computing*. Computational Methods in Applied Sciences. Springer International Publishing, Cham, Submitted April 2017
   T.K provided all demonstrations, provided most of final text.

4. Kirchdoerfer, T., and Ortiz, M. Data-driven computing in dynamics. *International Journal for Numerical Methods in Engineering* (Submitted June 2017)
   T.K Derived dynamics projection, implemented all code and method developement with significant contributions to writing of paper.

Contents

List of Figures

Chapter 1

Introduction

Significant text from this chapter is taken from [58].

1.1 Motivation

The computational sciences, as applied to physics and engineering problems, have always been concerned with using data inputs to provide solution results. Most of the methodologies that have been developed since the dawn of modern numerical analysis in the 1950's have been preoccupied with the discretization of space and time. Finite differences, finite elements, finite volumes, molecular dynamics, and mesh free methods are all examples of different ways of estimating solution fields. The constitutive data that gives rise to the predictive validity of these methods has been used to create models which are then embedded into the various solution methods. These models are designed to act as succinct summaries for, at times, complicated material responses for which summarization is a difficult task. Machine learning has been applied to automate this summarization process, but there remain real difficulties in using reduced forms to accurately reproduce complex phenomena. Primary among issues that restrict model quality is the need to characterize responses across regimes that are *data sparse*. Frequently, experimental data is restricted in its ability to collect the measurements needed to fully describe the domain of interest. In cases such as this, modeling is required to both summarize data and make use of inference to operate in under-sampled regimes.

At present, we are transitioning into an era of data generation and collection, where the availability of information sets that are *data rich* throughout the regimes of interest is becoming more common. Presently for problems in mechanics, complex sub-scale numerical models can be easily

exercised to systematically generate descriptive data sets of macroscopic field response. The problems for which experimental data is similarly descriptive remains comparably limited, however future experimental methods might eventually extend the range of experimentally data-characterized problems for mechanics. However, regardless of the data source, once a data set is such that data-inference is no longer required, empirical summaries are necessarily less rich than the data upon which they were based. In these circumstances, modeling then finds itself unable to take full advantage of the increasingly large data sets. Ultimately, the assumed properties of a model become a restriction on the ability of a calculation to converge to measured behavior. This lack of convergence then leads to unresolvable modeling errors, which ultimately influence the quality of the calculated solution fields. The question then becomes how to move scientific computing beyond the modeling paradigm and have it operate directly on the supplied data sets. In its most general form, *Data Science* is the extraction of *knowledge* from large volumes of unstructured data [8, 9, 7, 15]. It uses analytics, data management, statistics, and machine learning to derive mathematical models for subsequent use in decision making. Data Science already provides classification methods capable of processing source data directly into query answers in non-STEM problems, but no analogous method exists to perform scientific calculations. What remains then is a need for a method to use constitutive data to link the kinematic and kinetic laws properties which are at the core of any scientific calculation.

1.2 Thesis scope and overview

The work of this thesis is focused on developing a new scientific computational paradigm for Data Science known as *Data Driven Computing*. Previous works making use of Data Science in the service of scientific computing have focused on making use of traditional machine learning techniques to develop models for predicting new material properties, or generally automating the modeling process. Public materials databases also exist, but at present the contained data are themselves the parameterized outcomes of modeling processes. Data Driven Computing instead reformulates initial-boundary-value problems around data associations defined at integration points within the discretized domain. As a result, constitutive relations are explicitly defined by the source data associations. Material modelling empiricism, error, and uncertainty are then eliminated entirely with no loss of constitutive information.

What follows in the remainder of Chapter 1 first provides descriptions of the previous Data Science applications in Section 1.3 to provide contrast to the current work. Section 1.4 then transitions to the important commonalities of Data Driven Computing methods before Section 1.5 finishes with summarized descriptions of the new methods developed in Chapters 2-4. A distance-minimizing scheme is developed in Chapter 2, which is shown to converge for sequences of data sets which uniformly converge to a functional form. Quasistatic simulations then act as the primary demonstration case for Data Driven distance-minimization. These demonstrations are replicated and extended in Chapter 3 using a max-ent based clustering argument to thermalize the method within the context of an annealing schedule. This analytic extension of the distance-minimizing solvers developed in Chapter 2 makes this new class of data solvers robust in the presence of data sequences with persistent noise. Beyond the improvement of quasistatic response for noisy data sets, this thermalized extension proves especially valuable in the multi-step evolution of dynamics calculations explored in Chapter 4.

These beginning stages of development for Data Driven Computing demonstrate new possibilities for a young paradigm of computational science. Chapter 5 details some of the method extensions and applications that would naturally follow from this work, as well provide concluding remarks on the work completed. Within mechanics these new methods offer the potential to provide powerful scale-linking capacity to multi-scale analysis. Data Driven solvers move constitutive data away from being a loose substrate upon which analysis resides, into becoming a core constituent of the analytical process. This process provides the techniques with new forms of causality and convergence that will make their use foundational to any number of new strategies and foci in computational prediction.

1.3 State-of-the-art and differences with previous work

It is important to frame Data Driven Computing within the context of past and present efforts to automate and strengthen the connection between data and science. It is also particularly important to recognize the unique aspects of Data Driven Computing and how it differs from other efforts.

1.3.1 Material Informatics

There has been extensive previous work focusing on the application of Data Science and Analytics to material data sets. The field of Material Informatics [22, 23, 69, 81, 82, 80, 84, 83, 85, 26, 73, 31, 21, 51, 50, 42, 47, 49] uses data searching and sorting techniques to survey large material data sets. It also uses machine-learning regression [19, 93] and other techniques to identify patterns and correlations in the data for purposes of combinatorial materials design and selection. These approaches represent an application of standard sorting and statistical methods to material data sets. While efficient at looking up and sifting through large data sets, it is questionable that any real epistemic knowledge is generated by these methods. What is missing in Material Informatics is a direct use – and solution of – the field equations of physics as a means of constraining and ascertaining material behaviour. By way of contrast, such field equations play a prominent role in Data Driven Computing and make the approach *predictive*, and not just *postdictive*.

1.3.2 Material identification

There has also been extensive previous work concerned with the use of empirical data for parameter identification in prespecified material models, or for automating the calibration of the models. For instance, the Error-in-Constitutive-Equations (ECE) method is an inverse method for the identification of material parameters such as the Youngs modulus of an elastic material [41, 16, 35, 17, 20, 72, 27, 96, 10, 77, 68, 75]. While such approaches are efficient and reliable for their intended application, namely, the identification of material parameters, they are radically different from Data Driven Computing: material identification schemes aim to determine the parameters of a prespecified material law from experimental data; Data Driven Computing dispenses with material models altogether and uses material data directly in the formulation of initial-boundary-value problems and attendant calculations thereof. In particular, in Data Driven Computing no *a priori* assumptions are made regarding material behaviour, and the material data that is generated and used is *fundamental, unbiased, and model-independent*.

1.3.3 Data repositories

A number of repositories are presently in existence aimed at data-basing and disseminating material property data, e.g., [4, 2, 1, 3]. Some of these repositories receive extensive support from govern-

mental agencies and other sources. However, it is important to note that the existing material data repositories archive parametric data that is specific to the prespecified material models, which considerably limits their value and usefulness and puts them at variance with Data Driven Computing. For instance, OpenKIM [1] provides parametrizations of standard interatomic potentials, such as the embedded-atom method (EAM), for a wide range of materials systems. Evidently, such data is strongly biased by the assumption of a specific form of the interatomic potential. By way of sharp contrast, Data Driven Computing supports the development of data repositories that store fundamental, unbiased, model-independent material data only. Thus, suppose that the field equations of interest are the equations of molecular dynamics. In this case, the fundamental fields are the particle position and force fields, and the local states consist of atomic positions and corresponding forces over local clusters of atoms. It thus follows that, in this case, model-free unbiased data takes the form of local atomic positions and forces, determined, e.g., by means of first-principles quantum-mechanical calculations. For more complex mesoscopic systems, the role of mathematical analysis in determining what constitutes fundamental and unbiased material data – and what unit-cell problem determines the data – is of the essence.

In summary, the Data Driven Computing paradigm is unique in that it entirely bypasses any reference to preconceived empirical material models, i. e., it is *strictly model-free*, and incorporates directly – and without bias – into initial-boundary-value problems data that is *fundamental*, i. e., that lives in the natural phase space of the governing field equations.

1.4 Problem overview of Data Driven Computating

In this secton we discuss the primary aims and requirements of solution methods as they relate to the developement of solvers which are Data Driven. While there is significant variation in how Data Driven Computing can be affected with regards to solver methodologies and application constraints, there remain commonalites whose discussion would provide clarity in the following chapters. Subsection 1.4.1 provides a description of the general nature of constituive relations in scientific computing and articulates the problems created by using models to define these relations. This is followed by Subsection 1.4.2, which discuses how Data Driven solvers relax constitutive constraints between fields to accomodate the imposition of discrete points sets as constitutive relations. Finally Section 1.4.3 discusses the need for convergence demonstrations and their importance in underestanding the qualities of a given data solver.

1.4.1 Review of constitutive relations

A main task of scientific calculations is to resolve coupled field responses to boundary conditions. Constitutive relationships then define the nature of coupling between the related fields. In mechanics, the relations of interest are the extensive kinematic and kinetic work conjugate fields, e.g. ε and σ. The language here restricts itself to mechanics, but this focus exists as a special case of potential field theory through which electrostatics, diffusion and others present a similar need for constitutive definitions. Individually these fields must strictly satisfy *material independent* properties. Kinematic fields must satisfy compatibility, while kinetic fields conserve momentum to be consistent with known physical laws. The certainty with which such field constraints can be asserted stands in stark contrast to the *material dependent* constitutive model which typically relates the two fields. Such models must be informed by data, whose summarization into a model is typically performed using empirical fits which require sigficant knowledge and understanding of underlying phenomena. These model forms provide computational speed and the capacity for inference, but at the cost of a modeling error that influences computational conclusions in ways which are hard to characterize. Removing model characterization from the simulation process then allows for the removal of a potential source of simulation error and introduces opportunities for automation.

1.4.2 Develping relaxation schemes

To move beyond modeling, this thesis focuses upon the material data sets upon which such a models are based. We begin by first defining a finite point set, E, which exists in phase space Z, where $z = (\varepsilon, \sigma)$, and an example from small deformation mechanics would express the set as

$$E = ((\varepsilon_i, \sigma_i), i = 1, \cdots, N).$$

The discrete nature of the set would naturally confound constitutive strategies which rely upon making use of a characterized function form. If compatibility, equilibrium, and boundary conditions are represented by the constraint set C, a problem arises in the likely case where the combined constraints cannot be satisfied by couplings defined by the discrete data set, thus $E \cap C$ returns an empty set. Knowing this, we then seek a relaxation which continues to satisfy all the members of C, while minimizing deviations from E through direct data associations.

In exploring the need for a constitutive relaxation, it is helpful to first recall how constitutive relationships are affected in spacial discretization schemes. For any fixed domain, the corresponding fields are discretely sampled for enforcement at the integration points X . Any particular spacial discretization scheme then naturally defines these points as is deemed optimal by the associated method. This sampling, later written as $z \in C$, is more fully expressed as

$$z(x_i) = (\varepsilon(x_i), \sigma(x_i)),$$

where

$$X = (x_i, i = 1, \cdots, m),$$

$z \in Z$, and the constraint set C is satisfied, can imply that each of the fields is free to vary away from these points. As a result, since direct data associations $y \in E$ are discrete, it then naturally follows that the associations be defined at the similarly discrete integration points. This separation of the field values and data associations provides an elegant basis from which to formulate relaxed constitutive data methodologies. What remains to be explored, for any given implementation of these concepts, is the manner in which the relaxation is penalized to provide fidelity to the supplied data in the presence boundary constraints. Chapters 2 and 3 each discuss a different penalization

as a means of supplying different forms of data fidelity.

1.4.3 Demonstration requirements: Data Convergence

While relaxation and penalization provide the implementation details of how Data Driven solvers operate, these new methods require demonstrable forms of convergence which are different than is typically used in numerical analysis. Convergence is typically defined in terms of refinement metrics; however the dependence of data driven methods upon the data sets prevents such direct demonstrations. Because convergence cannot be guaranteed for an unknown sequence of data sets, Data Driven convergence demonstrations can only show solutions which are consistent with intrinsic features of the data, *should they exist*. This relaxed form of convergence demonstration is then implemented by embedding models within sequences of constructed data sets which are used as a means of testing Data Driven solvers for convergence. The manner in which this embedding process can be performed, so as to allow a given Data Driven Computing method to data-converge, then defines an important characteristic of the devised solver. Depending on the constitutive data set employed, understanding these solver properties would aid in the selection of the proper method.

1.5 Chapter overviews

What follows in this section is a summary of the major work topics of this thesis. To provide a greater consistency of discussion in these summaries, some of the specific examples shown here are modified versions of those seen in subsequent chapters. By providing adjacent chapter summaries here in Chapter 1, focus can be directed into distilling the top level concepts and relations that drive and connect the work in the remainder of the thesis.

1.5.1 Distance minimizing solutions

In order to add specifics to the relaxations discussed in Section 1.4.2, the question becomes *how* to relate $z \in C$ and $y \in E$ to best reflect E as a constitutive description. The simplest demonstration of Data Driven Computing, which is discussed in Chapter 2, identifies its measure of deviation as the distance between z and y in a defined metric space, $d(z, y)$. Thus is defined the equivalent minimization problems

$$\min_{z \in C} \min_{y \in E} d(z, y) = \min_{y \in E} \min_{z \in C} d(z, y).$$

This formulation then defines an optimal solution which would provide the field values z satisfying constraints C, and the data associations y in the material set E.

The Data Driven Computing scheme just outlined is illustrated in Figure 1.1 by means of the elementary example of a uniformly-deformed bar deforming under the action of a loading device of known behavior. As discussed, the constraint set C and the material data set E have an empty intersection. Through the selection of a relaxed minimization criteria, an ideal solution can be identified without the crutch of an assumed constitutive form. Beyond the need for solution results to be processed directly from a material data set, a Data Driven method needs to also be capable of exhibiting data-convergence. Such convergence is shown for a sequence of data sets E_i with corresponding size n_i, where $n_{i-1} > n_i$ if the solution converges as $i \to \infty$. For example, the constraint set and data set shown in Figure 1.1 make it apparent that under a distance minimizing scheme, data-convergence would be achieved if the data sets converge to a graph in phase space.

As a way to demonstrate the viability of distance minimizing Data Driven computation, Chapter 2 focuses on quasistatic test cases. While the chapter contains two working examples of such problems, the most thorough of these initial explorations targeted a small deformation, hyperelastic

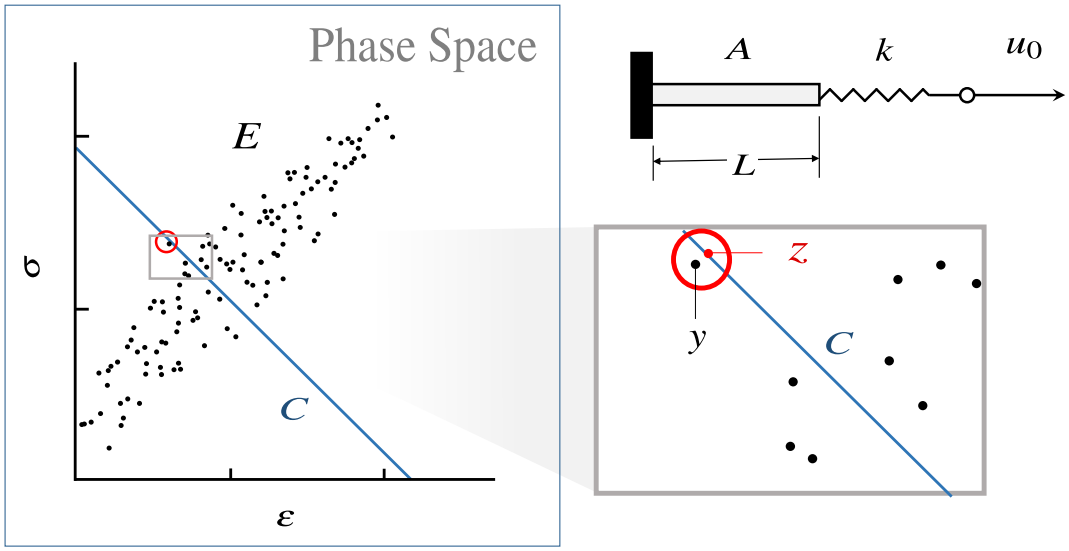


Figure 1.1: Bar loaded by soft device. The line C is the constraint set consistent with the applied displacement u_0. The material data set is the point set E. The distance minimizing, Data Driven solution is the red-encircled point y and projection z, which generate the smallest distance in the phase space.

truss with 1,048 degrees of freedom and a mix of specified boundary conditions. Figure 1.2 shows the geometry and boundary conditions of this test case, along with magnified deformation effects generated from the material data shown on the right. The diagram illustrates both how the constitutive relation of the bar is defined by a data set, and the specific nature of a data association. The distance function that defines the metric space for a stress-strain (σ,ε) phase space can be expressed as

$$d^2(z_1, z_2) = \frac{E}{2}(\varepsilon_2 - \varepsilon_1)^2 + \frac{1}{2E}(\sigma_2 - \sigma_1)^2,$$

where $z = (\varepsilon, \sigma)$ and E is a selected weighting modulus. The full optimization problem then becomes

$$\min_{z^{dat} \in E} \min \left(\sum_{e=1}^{n} w_e \left(\frac{E}{2}(B_e u - \varepsilon_e^{dat})^2 + \frac{1}{2E}(\sigma_e - \sigma_e^{dat})^2 \right) - \mu^T \left(\sum_{e=1}^{n} w_e B_e^T \sigma_e - f \right) \right),$$

where substitution is used to enforce compatibility, $\varepsilon = Bu$, and equilibrium is imposed in the final term using the Lagrange multipliers μ. This ordering of the function minimums represents the solution as one that minimizes the distance between the data association set and its projection onto

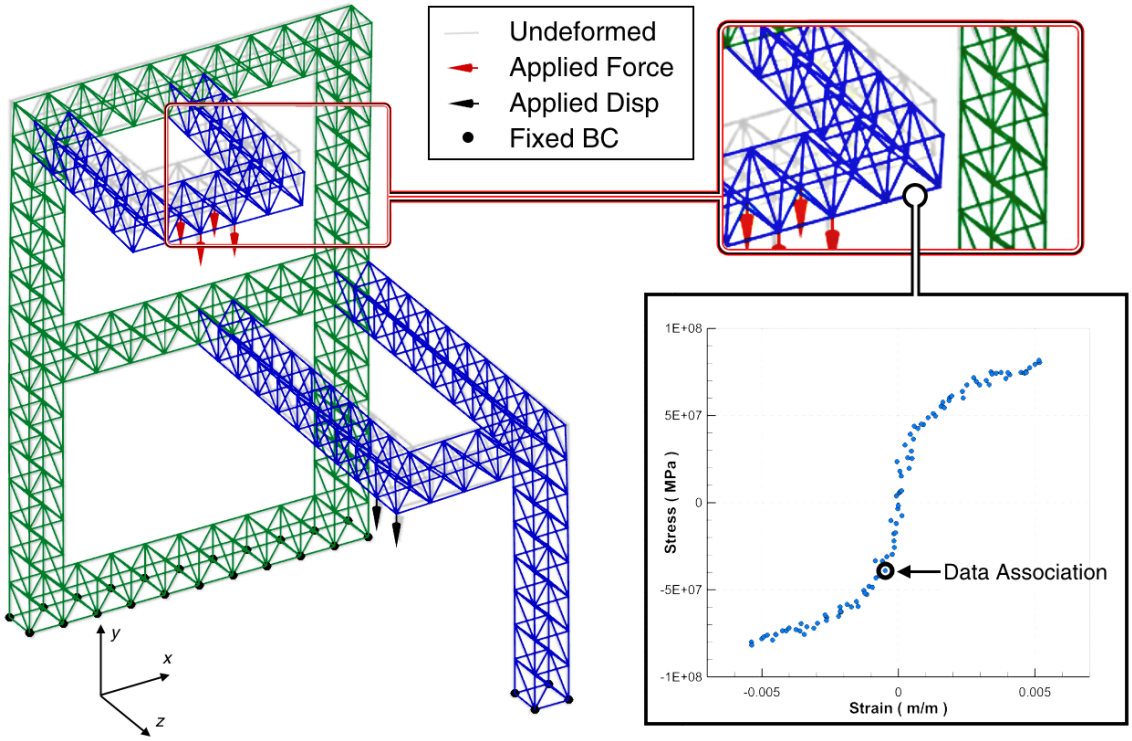


Figure 1.2: Static equilibrium of a three-dimensional truss in which the behavior of the material is known only through the data set shown.

the constraint set. Since the expression of all possible data associations includes all permutations of the bar-wise data assignments, the method cannot rigorously explore the entire material data space. Instead a fixed point iteration scheme is formulated to create a series of reductions in the minimizing function. This scheme, starting with an arbitrary assignment, projects the association set onto the constraint set. The projected solution is then used to perform a nearest neighbor search for new data associations. Termination then occors when all the data selections represent the closest point to their own projection.

Chapter 2 includes demonstrations with sufficient size and complexity (e.g., Figure 1.2) to show the viability of the method beyond the confines of simple examples. Numerical demonstrations of distance minimizing Data Driven calculations are seen to exhibit data-convergence when the data sets uniformly converge to a graph in the phase space. Solution costs are reasonable regarding both speed and complexity, while the material data set size most strongly impacts the computational expense of the described method. These initial demonstrations of efficacy provide initiation to the development of other Data Driven solvers as well as application of the method to new mechanics

problems.

1.5.2 Extensions to data sets with persistent noise

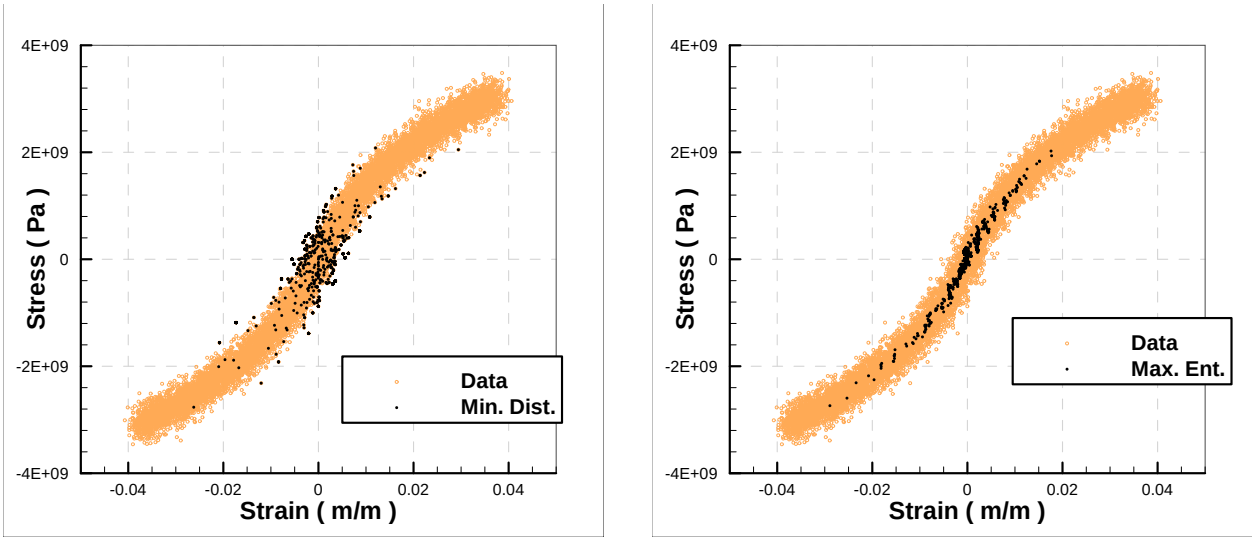
The initial work on Data Driven computing discussed in Chapter 2 focuses primarily on establishing and demonstrating a new class of Data Driven solvers. The distance-minimizing data solver discussed previously stands as an excellent vehicle for the exposition of this new class of methods, due in large part to the elegant simplicity of the distance-minimizing form. However, such solvers exhibit data-convergence for noisy sets only if the sequence of data sets converges to a graph in the phase space. The problem is adequately described by imagining what would happen to the data scenario pictured in Figure 1.1 if the data, instead of collapsing to a line, sees additions to the visualized set which are consistent with the pictured scatter. The points which best satisfy the constraint C would "hop" to a new point whenever one of the newly added points more closely approximates the constraint. Leaving aside the possibility of some distance ideal solution, $E \cap C$, this process would continue indefinitely, thus preventing the convergence to a final solution. In Chapter 3, data sets which contain a finite band are accommodated using a *probabilistic* solution strategy which arbitrates on the relevance and importance of different data points based on proximity. Distance-minimizing Data Driven solvers are incapable of data-convergence for banded data sets because they seek a single data member of E which most closely approximates the constraint set C. *Cluster analysis* provides a means of incorporating the influence of data neighborhoods to allow data-convergence in the presence of deeper samplings of fixed distributions.

Data Driven solvers for noisy data, developed in Chapter 3, employ cluster analysis so as to make a new kind of data driven solvers robust to outliers and is well suited to data sources with finite data bands. The foundations of cluster analysis have their roots in concepts provided by Information Theory, such as *maximum-entropy* estimation [52]. Specifically, we wish to quantify how well a point z in phase space is represented by a point z_i in a material data set $E = (z_1, \cdots, z_n)$. Equivalently, we wish to quantify the *relevance* of a point z_i in the material data set to a given point z in phase space. We measure the relevance of points z_i in the material data set by means of *weights* $p_i \in [0, 1]$ with the property: $\sum_{i=1}^{n} p_i = 1$. We wish the ranking by relevance of the material data points to be *unbiased*. It is known from Information Theory that the most unbiased distribution of weights is that which maximizes *Shannons information entropy* [90, 91, 89]. In addition, we wish

to accord points distant from z less weight than nearby points. These competing objectives can be combined by introducing a Pareto weight $\beta \geq 0$. The optimal and least-biased distribution is given by the Boltzmann distribution[89, 13]:

$$p_i = \frac{1}{Z} \exp\left(-\beta d^2(z, z_i)\right), \qquad Z = \sum_{i=1}^{n} \exp\left(-d^2(z, z_i)\right),$$

where the corresponding *max-ent* Data Driven solver now consists of minimizing the free energy $F(z) = -\log Z/\beta$ over the constraint set C. Making use of fixed point iteration, the method uses a field solution z to weight the summed data associations, which are then projected onto C to generate an update for z. We note that, since both methods make use of the same projection, the distance-minimizing Data Driven scheme is recovered in the limit of $\beta \to \infty$. For finite β, all points in the material data set influence the solution, but their corresponding weights diminish with distance to the trial solution z. The clustering scheme just described is in analogy to information-theoretical methods for reconstructing geometrical objects and functions from point data sets [13, 32].



(a) Truss data selections using distance minimizing method

(b) Truss data selections using entropy maximizing method

Figure 1.3: Minimum-distance and maximum-entropy Data Driven solvers point selections for a noisy 1000 point data set

Since the Pareto weight β is not based on any physical or data parameter, an annealing schedule is used to remove any direct solution dependency. The schedule is initialized with β chosen to

provide a convex free energy after which the method proceeds to raise β via a sequence of contracting local convexity estimates. A free parameter λ is introduced to control the speed of annealing, where solutions for a fixed data set converge under reductions to the annealing rate at significant increases to computational cost. Figure 1.3a shows the data selections made by the distance-minimizing method to solve the truss shown in Figure 1.2 from the pictured data set. In comparison, Figure 1.3b presents how the process of annealed data clustering produces selections along the center of the distribution for the same data set. Demonstrations of maximum entropy methods on the pictured static truss showed the method to data-converge faster than the equivalent distance-minimizing solver for banded data converging to a graph. Additionally, Chapter 1.3b shows the entropy maximizing method is also numerically shown to data-converge for material data sets sampled from a fixed, finitely banded, probability distribution in the phase space. While the demonstrations are computationally expensive, little effort is invested in computational efficiency and there exist several simple strategies that would greatly improve the algorithm performance. Chapter 3 ultimately presents a significant extension of Data Driven Computing to a much broader class of possible data problems.

1.5.3 Dynamics

Having worked to first extend Data Driven Computing to a broader class of data solvers, Chapter 4 implements a new type of constraint set, *implicit dynamics*. The dynamic extension of the static Data Driven solvers only requires substituting a discretized momentum balance for the static equilibrium equation used previously. Stationarity, instead of providing two uncoupled linear systems for u and μ, provides two linear systems which are coupled in both variables. It has been found that, with limited modification, both discussed methods admit the dynamic equilibrium constraints; however performance for max-ent solvers is generally better for any given data set. Numerical tests in Chapter 4 show convergence metrics which are consistent with results presented in the performance assessments seen in Chapter 2 and 3. Entropy-maximizing solvers provide better results in the presence of noisy data sets, the compounding error of multi-step solutions drives their solutions to dynamics calculations to be essential in the presence of a finitely banded data. To create a dynamics test case, the quasi-static case from Figure 1.2 is modified to include an oscillating base with no external force or displacement conditions. Making use of the data sets shown in Figure

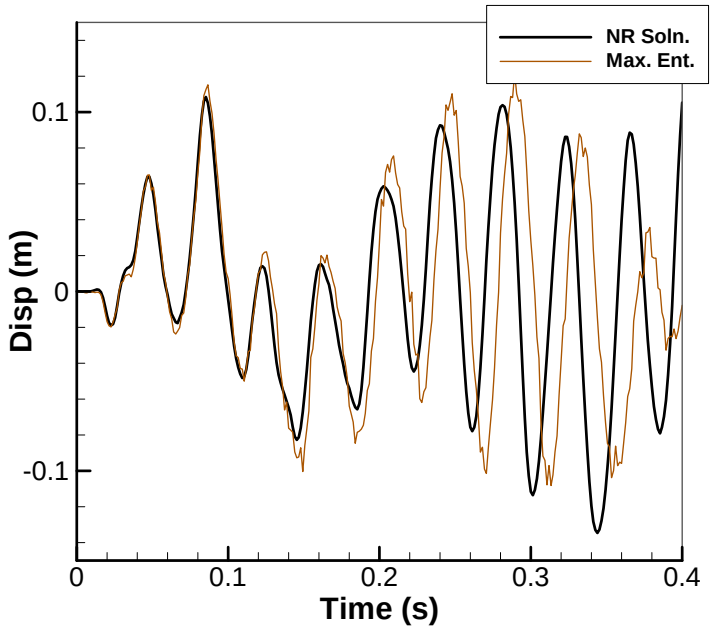


Figure 1.4: Dynamic x displacement response of node at top of a base-excited truss based on noisy constitutive data.

1.3, the deflection u_x at one of the former force application points is recorded. Figure 1.4 plots that deflection over 10 base oscillations compared to the results of a traditional non-linear implicit dynamics solver solution which made direct use of the model about which the pictured data is centered. Given the significant scatter of the shown input data, results show a remarkable correlation between the Data Driven and reference solutions.

Chapter 2

Distance Minimizing Data Driven Computing

An earlier version of this work is available as a preprint [54] and has since been published [55]. It is presented here with only small modifications.

2.1 Introduction

Boundary-value problems in science and engineering typically combine two types of equations: i) *conservation laws*, which derive from universal principles such as conservation of momentum or energy and are, therefore, uncertainty-free; and ii) *material laws*, formulated through physical modeling based on experimental observation and that are, therefore, empirical and uncertain. The prevailing classical computational paradigm has been to calibrate empirical material models using observational data and then use the calibrated material model in calculations. This process of modeling *a fortiriori* adds error and uncertainty to the solutions, especially in systems with high-dimensional phase spaces and complex behavior. This modeling error and uncertainty arises 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. Furthermore, often the models used to fit the data are *ad hoc*, without a clear basis in physics or a mathematical criterion for their selection, and thus the process of modeling is mired in empiricism and arbitrariness. Indeed, the entire process of empirical material modeling, and model validation thereof, is open-ended and no rigorous mathematical theory exists to date that makes it precise and quantitative.

Previous work has been carried out with a view to incorporating observational data into

boundary-value problem solution methodologies, but typically with the aim of augmenting or automating, rather than replacing, the use and generation of material models. Material informatics uses database techniques to first identify parameters of correlation and then use machine-learning regression techniques [19] to ultimately provide predictive quantitative models [6]. Principal-component analysis provides methods of dimensional reduction that allow such modeling techniques to be applied [31]. These approaches have been extended to the generation of multi-scale modeling correlations between macroscopic and microscopic constitutive properties [21, 46, 65, 48, 10].

These efforts may be understood as instances of *Data Science*, the extraction of "knowledge" from large volumes of unstructured data [9, 15]. Data science often requires sorting through big-data sets and extracting "insights" from these data. Data science uses data management, statistics and machine learning to derive mathematical models for subsequent use in decision making. Data Science currently influences primarily fields such as marketing, advertising, finance, social sciences, security, policy, medical informatics, whereas the full potential of Data Science as it relates to high-performance scientific computing is yet to be realized. Despite these limitations, reference to Data Science does effectively serve the purpose of bringing data and artificial intelligence considerations to the forefront.

In this work, we propose a new and different paradigm, which we refer to as *data-driven computing*, consisting of formulating calculations *directly* from experimental material data and pertinent essential constraints and conservation laws, thus bypassing the empirical material modeling step of conventional computing altogether. In this new computing paradigm, essential constraints and conservation laws such as compatibility and equilibrium remain unchanged, as do all the numerical schemes used in their discretization, such as finite elements, time-integrators, *et cetera*. Such conservation laws confer mathematical structure to the calculations, and this mathematical structure carries over to the present data-driven paradigm. However, in sharp contrast to conventional computing, in data-driven computing the experimental material-data points are used directly in calculations *in lieu* of an empirical material model. In this manner, material modeling empiricism, error, and uncertainty are eliminated entirely and no loss of experimental information is incurred. Specifically, data-driven solvers seek to assign to each material point the state from a prespecified data set that is closest to satisfying the conservation laws. Equivalently, data-driven solvers aim to find the state satisfying the conservation laws that is closest to the data set. The resulting

data-driven problem thus consists of the minimization of a distance function to the data set in phase space subject to the satisfaction of essential constraints and conservation laws.

We provide an efficient implementation of data-driven computing and demonstrate the practicality of the approach is demonstrated by means of two examples of application, namely, the static equilibrium of a nonlinear three-dimensional truss and linear elasticity. In these tests, the data-driven solvers exhibit good convergence properties both with respect to the number of data points and with regard to local data assignment. The variational structure of the data-driven problem also renders it amenable to analysis. We show that, as the data set approximates increasingly closely a classical material law in phase space, the data-driven solutions converge to the classical solution. We also illustrate the robustness of data-driven solvers with respect to spatial discretization. In particular, we show that the data-driven solutions of finite-element discretizations of linear elasticity converge jointly with respect to mesh size and approximation by the data set. The mathematical analysis is also suggestive of a number of generalizations and extensions of the data-driven computing paradigm.

2.2 Truss structures

We proceed to introduce and motivate the general approach with the aid of a simple stress-strain data set to solve a non-linear elastic truss problem. Trusses are assemblies of articulated bars that deform in uniaxial tension. Therefore, the material behavior of a bar is characterized by a particularly simple relation between uniaxial strain ε and uniaxial stress σ. We refer to the space of pairs (ε, σ) as *phase space*. We assume that the behavior of the material of each bar $e = 1, \ldots, m$, where m is the number of bars in the truss, is characterized by—possibly different—sets E_e of pairs (ε, σ), or *local states*. For instance, each point in the data set may correspond to, e. g., an experimental measurement, a subgrid multiscale calculation, or some other means of characterizing material behavior. A typical data set is notionally depicted in Fig. 2.1.

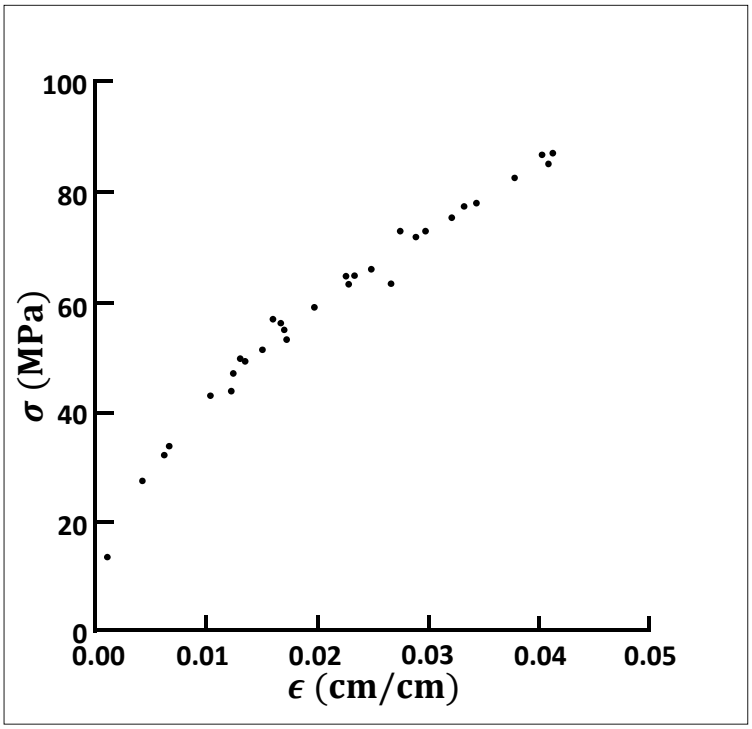


Figure 2.1: Typical material data set for truss bar.

2.2.1 Data-driven solver

For a given material data set, the proposed data-driven solvers seek to assign to each bar $e = 1, \ldots, m$ of the truss the best possible local state $(\varepsilon_e, \sigma_e)$ from the corresponding data set E_e, while simultaneously satisfying compatibility and equilibrium. We understand optimality of the local state in terms of an appropriate figure of merit that penalizes distance to the data set in phase space. For definiteness, we consider local penalty functions of the type

$$F_e(\varepsilon_e, \sigma_e) = \min_{(\varepsilon'_e, \sigma'_e) \in E_e} \Big(W_e(\varepsilon_e - \varepsilon'_e) + W_e^*(\sigma_e - \sigma'_e) \Big), \tag{2.1}$$

for each bar $e = 1, \ldots, m$ in the truss, with

$$W_e(\varepsilon_e) = \frac{1}{2} C_e \varepsilon_e^2 \quad \text{and} \quad W_e^*(\sigma_e) = \frac{1}{2} \frac{\sigma_e^2}{C_e} \tag{2.2}$$

and with the minimum taken over all local states $(\varepsilon'_e, \sigma'_e)$ in the local data set E_e. We may regard W_e and W_e^* as reference strain and complementary energy densities, respectively. We emphasize that

the functions W_e and W_e^* are introduced as part of the numerical scheme and need not represent any actual material behavior. In particular, the constant C_e is also numerical in nature and does not represent a material property.

Given a global state consisting of the collection of the local states $(\varepsilon_e, \sigma_e)$ of each one of its bars, the combined penalty function

$$F = \sum_{e=1}^{m} w_e F_e(\varepsilon_e, \sigma_e), \tag{2.3}$$

penalizing all local departures of the local states of the bars from their corresponding data sets. Here and subsequently, $w_e = A_e L_e$ denotes the volume of truss member e, with A_e its cross-sectional area and L_e its length. Therefore, the aim is to minimize F with respect to the global state $\{(\varepsilon, \sigma)\}$ subject to equilibrium and compatibility constraints. These aims lead to the constrained minimization problem

$$\text{Minimize:} \quad \sum_{e=1}^{m} w_e F_e(\varepsilon_e, \sigma_e), \tag{2.4a}$$

$$\text{subject to:} \quad \varepsilon_e = \sum_{i=1}^{n} B_{ei} u_i \quad \text{and} \quad \sum_{e=1}^{m} w_e B_{ei} \sigma_e = f_i, \tag{2.4b}$$

where $\{u_i,\ i = 1, \ldots, n\}$ is the array of displacement degrees of freedom, $\{f_i,\ i = 1, \ldots, n\}$ is the array of applied forces and the coefficients B_{ei} encode the connectivity and geometry of the truss.

The compatibility constraint can be enforced simply by expressing the strains in terms of displacements. The equilibrium constraint can be enforced by means of Lagrange multipliers, leading to the stationary problem

$$\delta \left(\sum_{e=1}^{m} w_e F_e\Big(\sum_{i=1}^{n} B_{ei} u_i, \sigma_e\Big) - \sum_{i=1}^{N} \Big(\sum_{e=1}^{m} w_e B_{ei} \sigma_e - f_i \Big) \eta_i \right) = 0. \tag{2.5}$$

Taking all possible variations, we obtain

$$\delta u_i \Rightarrow \sum_{e=1}^{m} w_e C_e \Big(\sum_{j=1}^{n} B_{ej} u_j - \varepsilon_e^* \Big) B_{ei} = 0, \tag{2.6a}$$

$$\delta \sigma_e \Rightarrow \frac{1}{C_e} (\sigma_e - \sigma_e^*) = \sum_{i=1}^{n} B_{ei} \eta_i, \tag{2.6b}$$

$$\delta \eta_i \Rightarrow \sum_{e=1}^{m} w_e B_{ei} \sigma_e = f_i, \tag{2.6c}$$

where $(\varepsilon_e^*, \sigma_e^*)$ denote (unknown) optimal data points for each one of the bars, i. e., data points such that

$$F_e(\sum_{i=1}^{n} B_{ei} u_i, \sigma_e) = W_e(\sum_{i=1}^{n} B_{ei} u_i - \varepsilon_e^*) + W_e^*(\sigma_e - \sigma_e^*) \tag{2.7}$$

or

$$W_e(\sum_{i=1}^{n} B_{ei} u_i - \varepsilon_e^*) + W_e^*(\sigma_e - \sigma_e^*) \leq W_e(\sum_{i=1}^{n} B_{ei} u_i - \varepsilon_e') + W_e^*(\sigma_e - \sigma_e') \tag{2.8}$$

for all data points $(\varepsilon_e', \sigma_e')$ in the local data set E_e. Once all optimal data points are determined, eqs. (2.6) define a system of linear equations for the nodal displacements, the local stresses and the Lagrange multipliers. A straightforward manipulation of these equations renders them in the equivalent form

$$\sum_{j=1}^{n} \left(\sum_{e=1}^{m} w_e C_e B_{ej} B_{ei} \right) u_j = \sum_{e=1}^{m} w_e C_e \varepsilon_e^* B_{ei}, \tag{2.9a}$$

$$\sum_{j=1}^{n} \left(\sum_{e=1}^{m} w_e C_e B_{ei} B_{ej} \right) \eta_j = f_i - \sum_{e=1}^{m} w_e B_{ei} \sigma_e^*. \tag{2.9b}$$

We recognize in these equations two standard linear-elastic truss-equilibrium problems with identical stiffness matrix corresponding to the reference linear truss defined by W_e and W_e^*, $e = 1, \ldots, m$. The displacement problem (2.9a) is driven by the optimal local strains, whereas the Lagrange multiplier problem (2.9b) is driven by the out-of-balance forces attendant to the optimal local stresses.

It remains to determine the optimal local data points, i. e., the stress and strain pairs $(\varepsilon_e^*, \sigma_e^*)$ in the local data sets E_e that result in the closest possible satisfaction of compatibility and equilibrium. The determination of the optimal local data points can be effected iteratively. Initially, all bars in the truss are assigned random points $(\varepsilon_e^{*(0)}, \sigma_e^{*(0)})$ from the corresponding local data sets E_e. The

displacements $u_i^{(0)}$ and Lagrange multipliers $\eta_i^{(0)}$ are then computed by solving (2.9) and the stresses $\sigma_e^{(0)}$ are evaluated from (2.6b). The next local data assignment is then effected by determining, for every member in the truss, the data points $(\varepsilon_e^{*(1)}, \sigma_e^{*(1)})$ in E_e that are optimal with respect to the local state $(\varepsilon_e^{(0)}, \sigma_e^{(0)})$, i. e., such that

$$W_e(\varepsilon_e^{(0)} - \varepsilon_e^{*(1)}) + W_e^*(\sigma_e^{(0)} - \sigma_e^{*(1)}) \leq W_e(\varepsilon_e^{(0)} - \varepsilon'_e) + W_e^*(\sigma_e^{(0)} - \sigma'_e) \tag{2.10}$$

for all data points $(\varepsilon'_e, \sigma'_e)$ in the local data set E_e. This operation entails simple local searches in phase space. The iteration then proceeds by recursion and terminates when the local data assignments effect no change. A detailed flowchart of the data-driven solver is listed in Algorithm 1.

Algorithm 1 Data-driven solver

Require: Local data sets E_e, B_e-matrices, $e = 1, \ldots, m$. Applied loads f_i, $i = 1, \ldots, n$.

i) Set $k = 0$. Initial local data assignment:

for all $e = 1, \ldots, m$ **do**

Choose $(\varepsilon_e^{*(0)}, \sigma_e^{*(0)})$ randomly from E_e

end for

ii) Solve:

$$\sum_{j=1}^{n} \left(\sum_{e=1}^{m} w_e C_e B_{ej} B_{ei} \right) u_j^{(k)} = \sum_{e=1}^{m} w_e C_e \varepsilon_e^{*(k)} B_{ei}, \tag{2.11a}$$

$$\sum_{j=1}^{n} \left(\sum_{e=1}^{m} w_e C_e B_{ei} B_{ej} \right) \eta_j^{(k)} = f_i - \sum_{e=1}^{m} w_e B_{ei} \sigma_e^{*(k)}, \tag{2.11b}$$

for $u_i^{(k)}$ and $\eta_i^{(k)}$, $i = 1, \ldots, n$.

iii) Compute local states:

for all $e = 1, \ldots, m$ **do**

$$\varepsilon_e^{(k)} = \sum_{i=1}^{n} B_{ei} u_i^{(k)}, \qquad \sigma_e^{(k)} = \sigma_e^{*(k)} + C_e \sum_{i=1}^{n} B_{ei} \eta_i^{(k)} \tag{2.12}$$

end for

iv) Local state assignment:

for all $e = 1, \ldots, m$ **do**

Choose $(\varepsilon_e^{*(k+1)}, \sigma_e^{*(k+1)})$ closest to $(\varepsilon_e^{(k)}, \sigma_e^{(k)})$ in E_e.

end for

v) Test for convergence:

if $(\varepsilon_e^{*(k+1)}, \sigma_e^{*(k+1)}) = (\varepsilon_e^{*(k)}, \sigma_e^{*(k)})$ for all $e = 1, \ldots, m$, **then**

v.a) $u_i = u_i^{(k)}$, $i = 1, \ldots, n$.

v.b) $(\varepsilon_e, \sigma_e) = (\varepsilon_e^{(k)}, \sigma_e^{(k)})$, $e = 1, \ldots, m$.

v.c) **exit**.

else

$k \leftarrow k + 1$, **goto** (ii).

end if

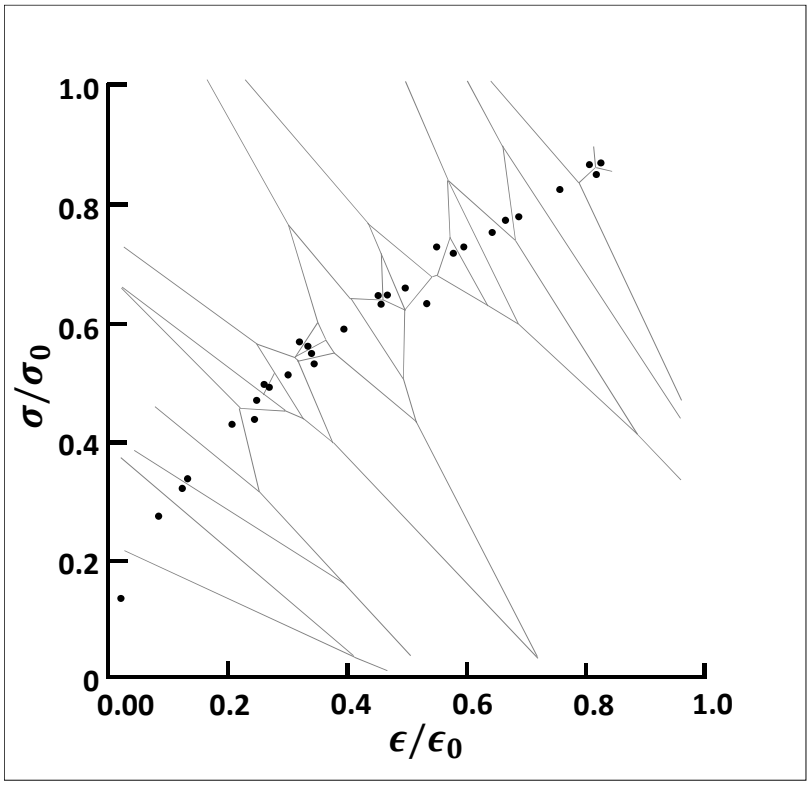


Figure 2.2: Voronoi tessellation of a data set.

The geometry of the local data assignment (2.10) is illustrated in Fig. 2.2. Thus, given a trial local state $(\varepsilon_e^{(k)}, \sigma_e^{(k)})$ of bar e, corresponding to the kth iteration of the solver, the next data point $(\varepsilon_e^{*(k+1)}, \sigma_e^{*(k+1)})$ assigned to the bar is the point in E_e that is closest to $(\varepsilon_e^{(k)}, \sigma_e^{(k)})$ in the norm

$$\|(\varepsilon_e, \sigma_e)\|_e = \Big(W_e(\varepsilon_e) + W_e^*(\sigma_e)\Big)^{1/2}. \tag{2.13}$$

This is precisely the data point $(\varepsilon_e^{*(k+1)}, \sigma_e^{*(k+1)})$ in E_e whose Voronoi cell contains $(\varepsilon_e^{(k)}, \sigma_e^{(k)})$. Thus, the penalty function (2.1) or, equivalently, the norm (2.13) divides the phase space into cells according to the Voronoi tessellation of E_e. Each cell in that tessellation may be regarded as the 'domain of influence' of the corresponding data point. The local state assignment then simply assigns material points according to their domain of influence and the iteration terminates when the local states of all bars lie within the domain of influence of the corresponding data points assigned to the bars.

2.2.2 Numerical analysis of convergence

A central question to be ascertained concerns the convergence of data-driven solvers with respect to the data set. Specifically, suppose that the materials in the truss obey a well-defined constitutive law in the form of a graph, or stress-strain curve, in (ε, σ)-phase space. Then, we expect the data-driven solutions to converge to the classical solution when the data sets approximate the stress-strain curve increasingly closely, in some appropriate sense to be made precise subsequently.

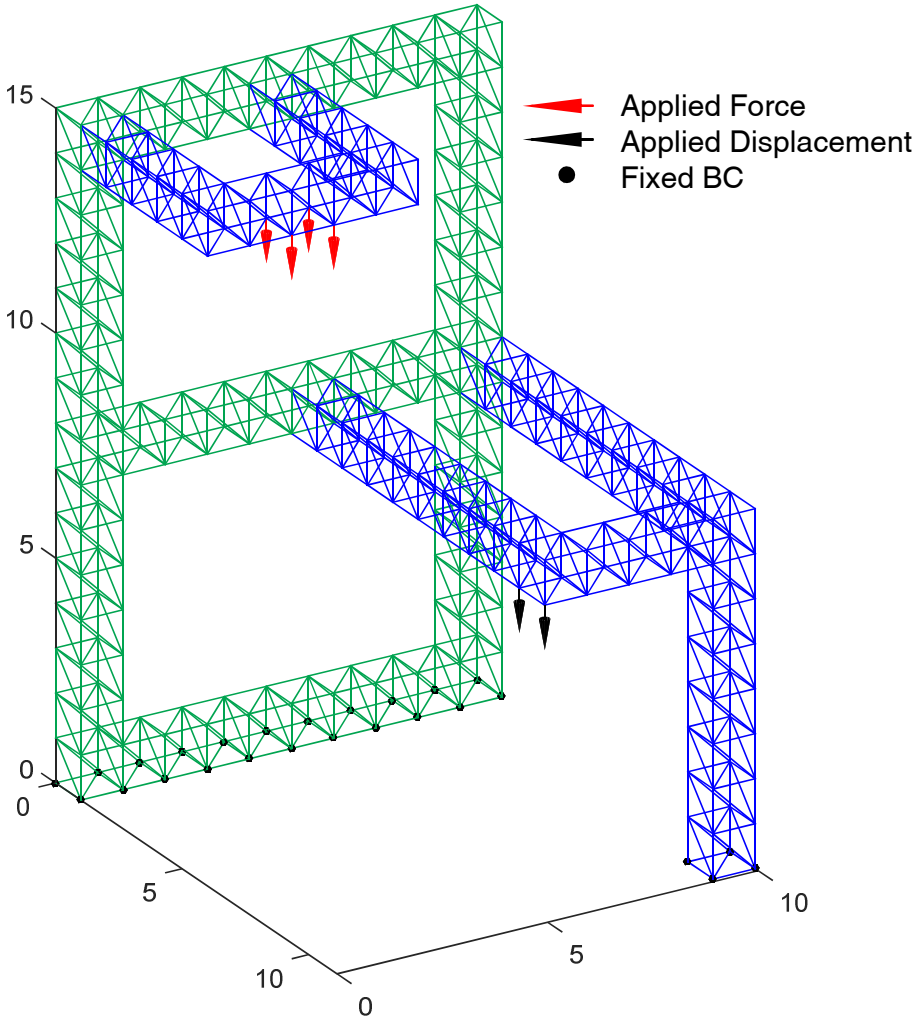


Figure 2.3: Model problem geometry with boundary conditions.

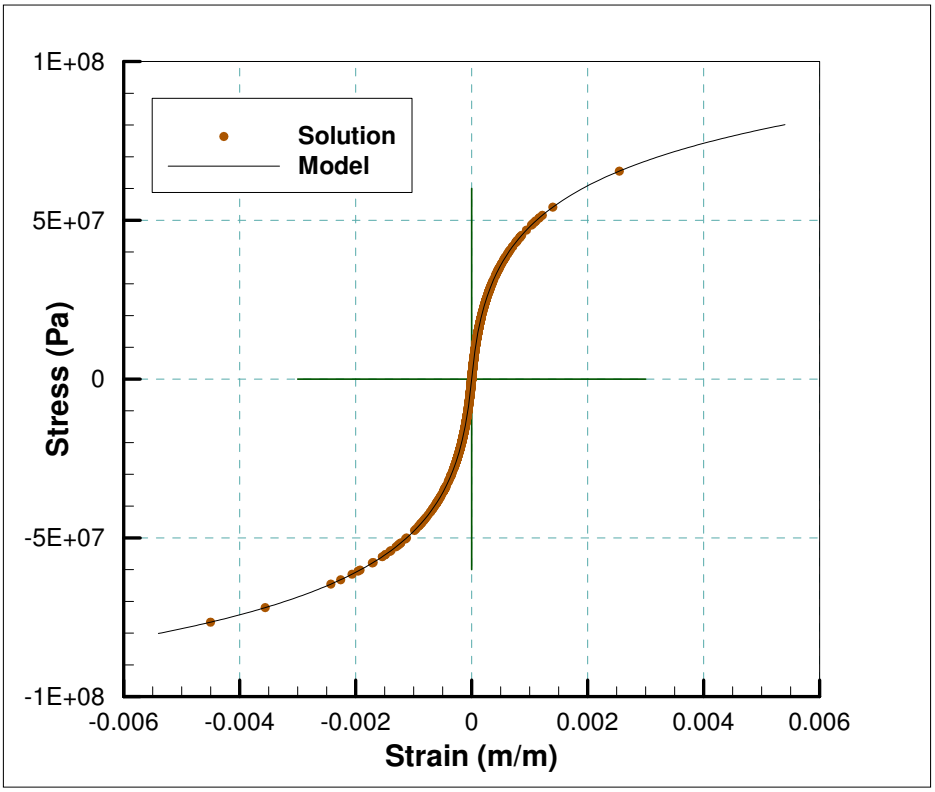


Figure 2.4: Material model with reference solution values superimposed.

In this section, we exhibit this convergence property in a specific example of application. Fig. 2.3 shows the geometry, boundary conditions and applied loads on a truss containing 1,048 degrees of freedom. The truss undergoes small deformations and the material in all bars obeys the nonlinear non-linear elastic law shown in Fig. 2.4. A Newton-Raphson solver is used to calculate the reference solution. The reference solution values thus obtained are plotted on the constitutive stress-strain curve to show the significant amount of non-linearity seen in the solution.

Suppose that, in actual practice, the stress-strain curve in Fig. 2.4 is not known exactly but instead sampled by means of a finite collection of points, or data sets. We begin by considering a sequence (E_k) of increasingly fine data sets consisting of points on the stress-strain curve at uniform distances $\rho_k \downarrow 0$, with distance defined in the sense of the norm (2.13).

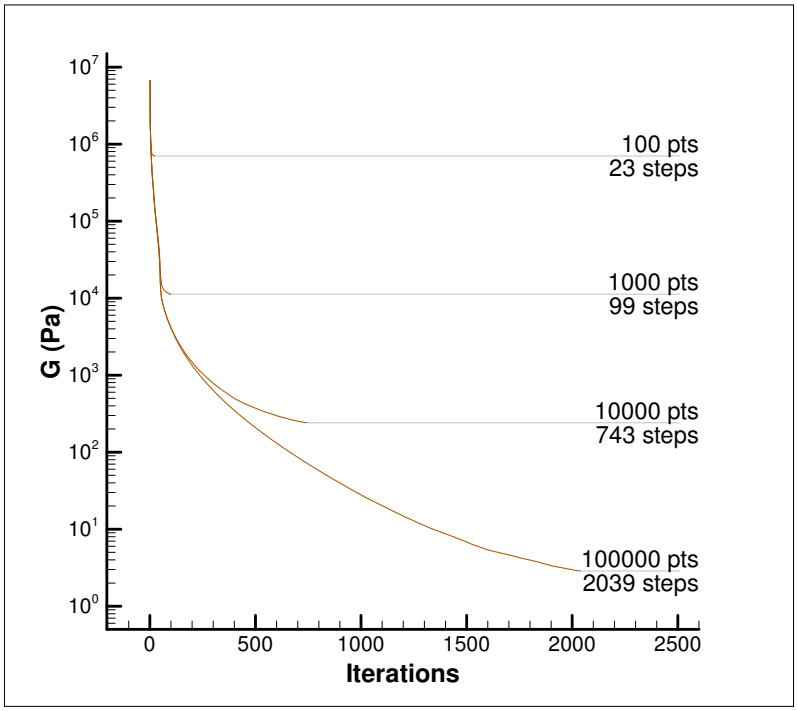


Figure 2.5: Convergence of the local data-assignment iteration.

The convergence of the local data assignment iteration is shown in Fig. 2.5 for data sets of sizes 10^2, 10^3, 10^4, and 10^5. In all cases, the initial local data assignment is random and convergence is monitored in terms of the penalty function F, eq. (2.3). We note that the problem of assigning data points optimally to each bar of the truss is of combinatorial complexity. Therefore, it is remarkable that the local data assignment iterations converges after a relatively small number of steps. As expected, the number of iterations to convergence increases with the size of the data set. However, it bears emphasis that each local data assignment iteration entails a linear solve corresponding to the linear comparison truss. The matrix of the system of equations, or stiffness matrix, can be factorized once and for all at the start of the iteration, and subsequent iterations require inexpensive back-substitutes only.

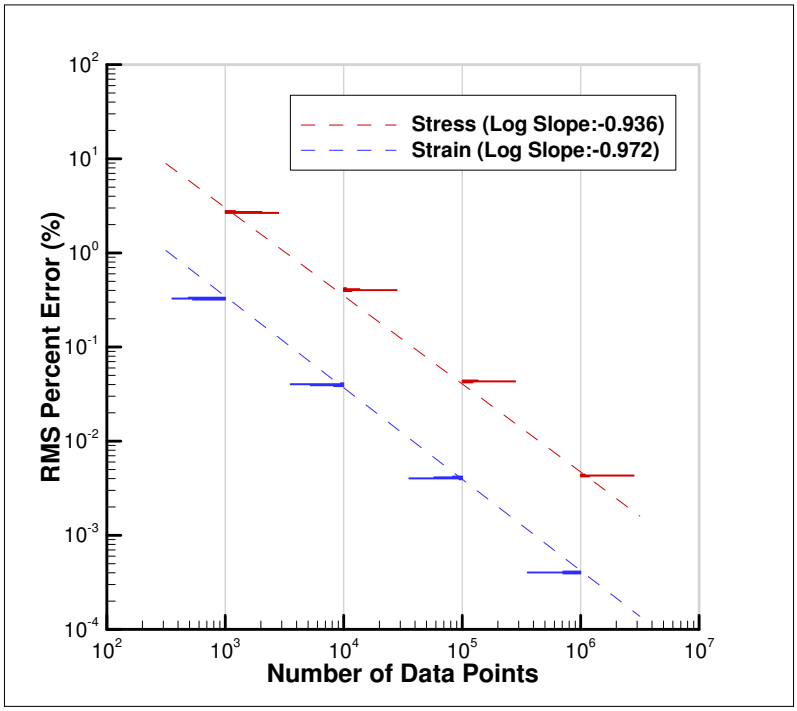


Figure 2.6: Convergence of strain and stress root-mean-square errors with number of sampling points. Histograms correspond to 30 different initial random assignments of data points to the truss members.

Next, we turn to the question of convergence with respect to the number of data points. For definiteness, we monitor the convergence of the resulting sequence of data-driven solutions to the reference solution in the sense of the normalized percent root-mean-square stress and strain errors

$$\varepsilon_{(\%\mathrm{RMS})} = \frac{1}{\varepsilon^{\mathrm{ref}}_{\max}} \left(\frac{\sum_{e=1}^{m} w_e (\varepsilon_e - \varepsilon^{\mathrm{ref}}_e)^2}{m} \right)^{1/2}, \tag{2.14a}$$

$$\sigma_{(\%\mathrm{RMS})} = \frac{1}{\sigma^{\mathrm{ref}}_{\max}} \left(\frac{\sum_{e=1}^{m} w_e (\sigma_e - \sigma^{\mathrm{ref}}_e)^2}{m} \right)^{1/2}, \tag{2.14b}$$

respectively, where the maximum $(\varepsilon^{\mathrm{ref}}_e, \sigma^{\mathrm{ref}}_e)$, $e = 1, \ldots, m$ are the strains and stresses corresponding to the reference solution and $(\varepsilon^{\mathrm{ref}}_{\max}, \sigma^{\mathrm{ref}}_{\max})$ are the corresponding maximum values.

Fig. 2.6 shows convergence plots of the strain and stress root-mean-square errors with number of sampling points. As may be observed from the figure, the convergence is close to linear in both strains and stresses, which verifies the convergence of the method as the data set approaches the presumed exact model. We recall that the data assignment algorithm 2.11 starts by randomly

assigning data points to the truss members. Evidently, the subsequent iteration depends on this initial choice. In order to demonstrate insensitivity to such initialization, convergence plots for 30 initial random assignments are shown in Fig. 2.6 and the resulting errors are binned into histograms. The tightness of these histograms verifies the robustness of the iteration with respect to the initial data point selection.

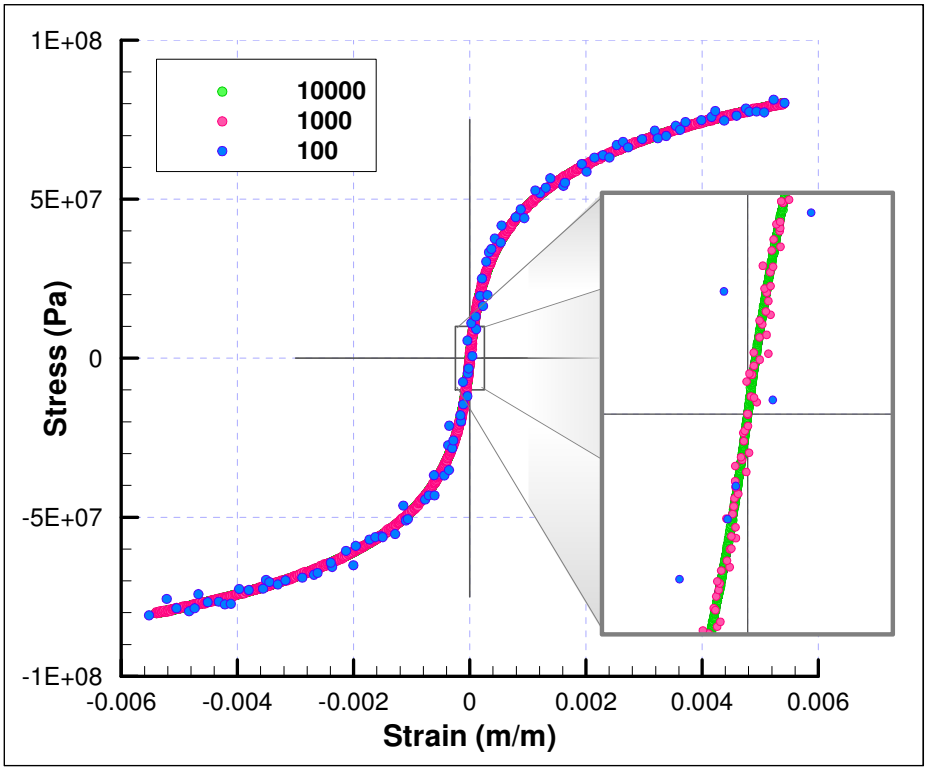


Figure 2.7: Typical data set with Gaussian random noise.

Next, we revisit the question of convergence with respect to the number of data points when the data set is noisy, i. e., when it does not sample the limit stress-strain curve but is offset from the curve with some probability. In this case, the data sets converge to the exact stress-strain curve as sets, in a manner to be made precise subsequently. In calculations we specifically begin by sampling the limit stress-stain curve at uniform distances $\rho_k \downarrow 0$, as in the preceding test cases, but then add Gaussian noise to the data points of variance ρ_k. A typical data set is shown in Fig. 2.7 by way of illustration. Convergence plots corresponding to 100 data sets are shown in Fig. 2.8. As may be seen from the figure, convergence is achieved with increasing number of points, albeit the convergence rate of roughly $1/2$ is lower than the convergence rate in the case of noiseless data.

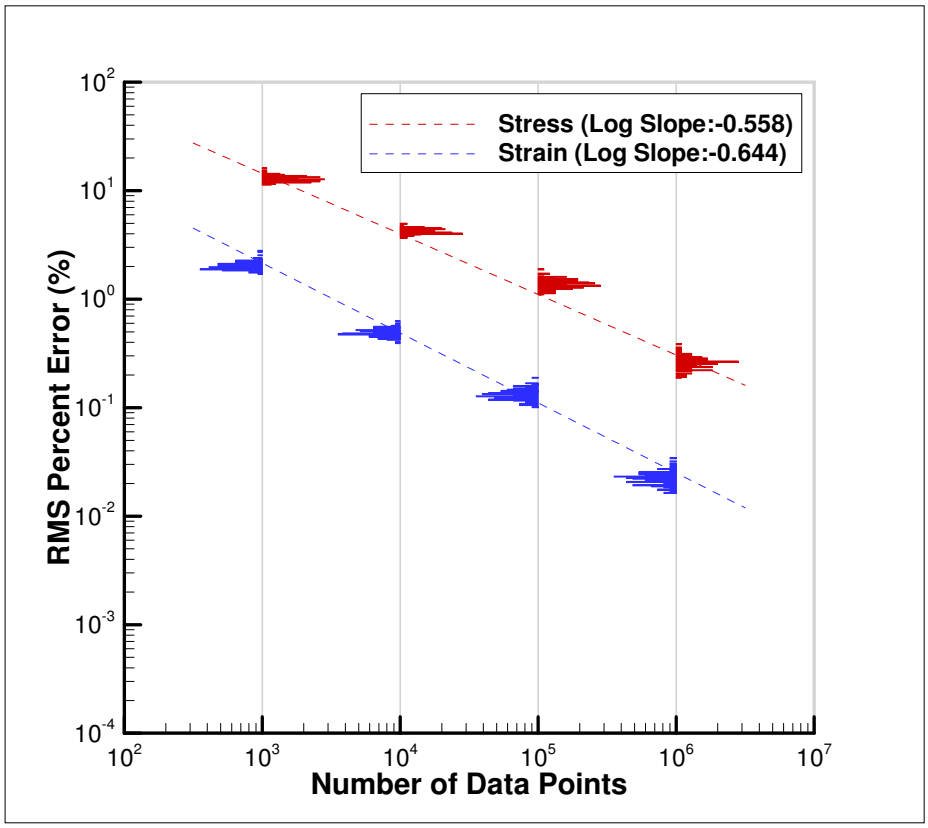


Figure 2.8: Convergence of strain and stress root-mean-square errors with number of sampling points and data sets with Gaussian noise. Histograms correspond to 100 data sets.

Finally, we examine the question of sample quality, i. e., the ability of a given data set to sample closely all the local states covered by the solution. Fig. 2.9 shows the distribution of the values of the local penalty function F_e, eq. (2.1) corresponding to data sets of sizes 10^2, 10^3, 10^4, 10^5. We recall that the value of the function F_e provides a measure of the distance of the local state $(\varepsilon_e, \sigma_e)$ to the data set. As may be seen from the figure, F_e tends to decrease with the number of sampling points, as expected. However, for every data-set size there remains a certain spread in the values of F_e, indicating that the states of certain truss members are better sampled by the data set than others. Specifically, truss members for which no data point lies close to their states result in high values of F_e, indicative of poor coverage by the data set. This analysis of the local values F_e of the penalty function suggests a criterion for improving data sets adaptively so as to improve their quality vis a vis a particular application. Evidently, the optimal strategy is to target for further testing the region of phase space corresponding to the truss members with highest values

of F_e. In particular, outliers, or truss members with states lying far from the data set, are targeted for further testing. In this manner, the data set is adaptively expanded so as to provide the best possible coverage of the distribution of local states corresponding to a particular application.

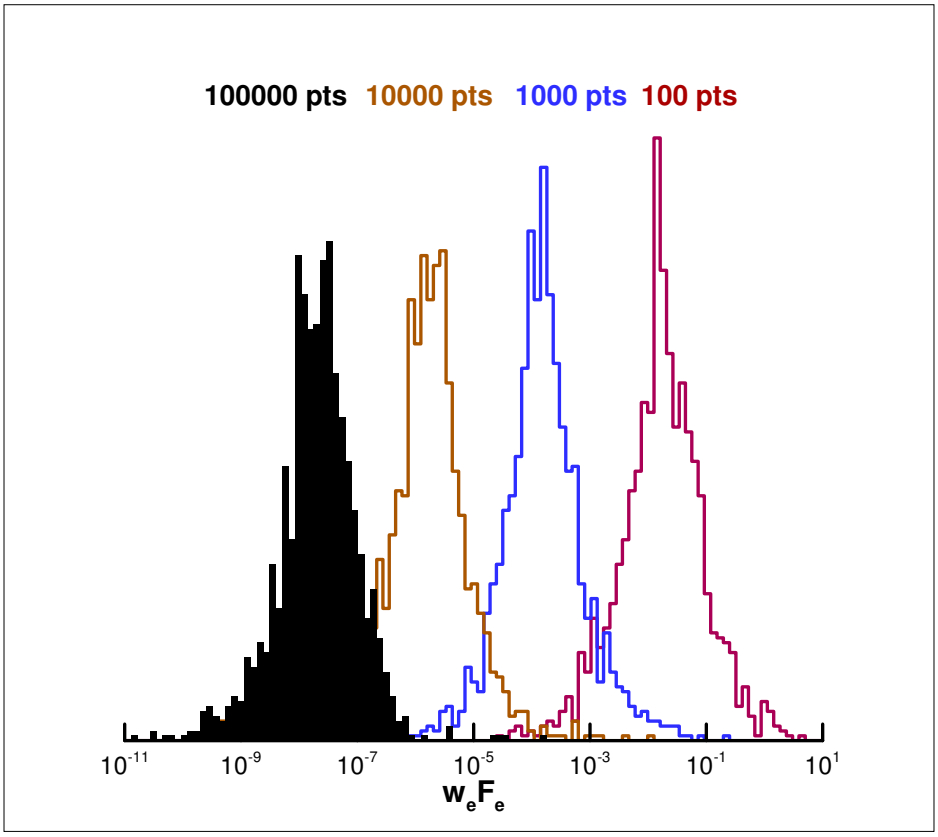


Figure 2.9: Distribution of values of local penalty functions $F_e(\varepsilon, \sigma)$ for converged data-driven solution.

2.3 Linear elasticity

As a second motivational example of application of the data-driven paradigm, we consider three-dimensional linear elasticity. In this case, the local phase space of the material consists of pairs $(\boldsymbol{\epsilon}, \boldsymbol{\sigma})$ of strain and stress, respectively. Since both stresses and strains are symmetric tensors, it follows that the corresponding phase space is twelve-dimensional. This dimensionality is high enough to start raising questions regarding material sampling and material-data coverage of the relevant region of phase space. An additional issue that is raised by linear elasticity concerns the infinite-dimensional character of the solution space. Thus, even if the problem is rendered finite-

dimensional by recourse to spatial discretization, the question of convergence with respect to mesh size must necessarily be elucidated within an appropriate functional framework. In this section, we extend the truss data-driven solver to linear elasticity and address the issue of data sampling in high dimensions by exploiting material and geometrical symmetry in the problem. Finally, we address the question of convergence of the finite-element discretized data-driven solver with respect to mesh size.

2.3.1 Data-driven solver

We consider a finite-element model of a non-linear elastic solid in the linearized kinematics approximation. The material behavior of a solid is characterized by a relation between the strain tensor $\boldsymbol{\epsilon}$ and the stress tensor $\boldsymbol{\sigma}$. We refer to the space of pairs $(\boldsymbol{\epsilon}, \boldsymbol{\sigma})$ as *phase space*. We assume that the behavior of the material or integration points in the model is characterized by—possibly different—sets E_e of pairs $(\boldsymbol{\epsilon}, \boldsymbol{\sigma})$, or *local states*, where $e = 1, \ldots, m$ labels the material points and m is the number of material points in the finite-element model.

We consider local penalty functions of the type

$$F_e(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e) = \min_{(\boldsymbol{\epsilon}'_e, \boldsymbol{\sigma}'_e) \in E_e} \Big(W_e(\boldsymbol{\epsilon}_e - \boldsymbol{\epsilon}'_e) + W_e^*(\boldsymbol{\sigma}_e - \boldsymbol{\sigma}'_e) \Big), \tag{2.15}$$

for each integration point $e = 1, \ldots, m$ in the solid, with

$$W_e(\boldsymbol{\epsilon}_e) = \frac{1}{2}\lambda(\mathrm{tr}\boldsymbol{\epsilon}_e)^2 + \mu\boldsymbol{\epsilon}_e \cdot \boldsymbol{\epsilon}_e \equiv \mathbb{C}_e\boldsymbol{\epsilon}_e \cdot \boldsymbol{\epsilon}_e, \tag{2.16a}$$

$$W_e^*(\boldsymbol{\sigma}_e) = \frac{1}{4\mu}\boldsymbol{\sigma}_e \cdot \boldsymbol{\sigma}_e - \frac{1}{4\mu}\frac{\lambda}{3\lambda + 2\mu}(\mathrm{tr}\boldsymbol{\sigma}_e)^2 \equiv \mathbb{C}_e^{-1}\boldsymbol{\sigma}_e \cdot \boldsymbol{\sigma}_e, \tag{2.16b}$$

with the minimum taken over all local states $(\boldsymbol{\epsilon}'_e, \boldsymbol{\sigma}'_e)$ in the local data set E_e. We may regard W_e and W_e^* are reference strain and complementary energy densities, respectively.

Given a global state consisting of a collection of local states $(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e)$ at each material point, we define a global penalty function as

$$F = \sum_{e=1}^{m} w_e F_e(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e), \tag{2.17}$$

w_e are quadrature or integration weights. This function penalizes jointly all departures of local

states from their corresponding data sets. The data-driven problem is to minimize F with respect to the global state $\{(\boldsymbol{\epsilon}, \boldsymbol{\sigma})\}$ subject to equilibrium and compatibility constraints, namely,

$$\text{Minimize:} \quad \sum_{e=1}^{m} w_e F_e(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e), \tag{2.18a}$$

$$\text{subject to:} \quad \boldsymbol{\epsilon}_e = \sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a \quad \text{and} \quad \sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \boldsymbol{\sigma}_e = \boldsymbol{f}_a, \tag{2.18b}$$

where $\{\boldsymbol{u}_a,\ a = 1, \ldots, n\}$ is the array of nodal displacements, $\{\boldsymbol{f}_a,\ a = 1, \ldots, n\}$ is the array of applied nodal forces, n is the number of nodes and the coefficients $\boldsymbol{B}_{ea}$ encode the connectivity and geometry of the solid.

As in the data-driven truss problem, the compatibility constraint can be enforced simply by expressing the strains in terms of displacements. The equilibrium constraint can in turn be enforced by means of Lagrange multipliers, resulting in the stationary problem

$$\delta \left(\sum_{e=1}^{m} w_e F_e(\sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a, \boldsymbol{\sigma}_e) - \sum_{a=1}^{N} \Big(\sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \boldsymbol{\sigma}_e - \boldsymbol{f}_a \Big) \boldsymbol{\eta}_a \right) = 0. \tag{2.19}$$

Taking all possible variations, we obtain the system of Euler-Lagrange equations

$$\delta \boldsymbol{u}_a \Rightarrow \sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \mathbb{C}_e \Big(\sum_{b=1}^{n} \boldsymbol{B}_{eb} \boldsymbol{u}_b - \boldsymbol{\epsilon}_e^* \Big) = 0, \tag{2.20a}$$

$$\delta \boldsymbol{\sigma}_e \Rightarrow \mathbb{C}_e^{-1} (\boldsymbol{\sigma}_e - \boldsymbol{\sigma}_e^*) = \sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{\eta}_a, \tag{2.20b}$$

$$\delta \boldsymbol{\eta}_a \Rightarrow \sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \boldsymbol{\sigma}_e = \boldsymbol{f}_a, \tag{2.20c}$$

where $(\boldsymbol{\epsilon}_e^*, \boldsymbol{\sigma}_e^*)$ denote the unknown optimal data points at material point e, i. e., the data point such that

$$F_e(\sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a, \boldsymbol{\sigma}_e) = W_e(\sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a - \boldsymbol{\epsilon}_e^*) + W_e^*(\boldsymbol{\sigma}_e - \boldsymbol{\sigma}_e^*), \tag{2.21}$$

or

$$W_e(\sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a - \boldsymbol{\epsilon}_e^*) + W_e^*(\boldsymbol{\sigma}_e - \boldsymbol{\sigma}_e^*) \leq W_e(\sum_{a=1}^{n} \boldsymbol{B}_{ea} \boldsymbol{u}_a - \boldsymbol{\epsilon}_e') + W_e^*(\boldsymbol{\sigma}_e - \boldsymbol{\sigma}_e'), \tag{2.22}$$

for all data points $(\boldsymbol{\epsilon}_e', \boldsymbol{\sigma}_e')$ in the local data set E_e. Once all optimal data points are determined,

eqs. (2.20) define a system of linear equations for the nodal displacements, the local stresses and the Lagrange multipliers. As in the data-driven truss problem, these equations can be rendered in the equivalent form

$$\sum_{b=1}^{n}\left(\sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \mathbb{C}_e \boldsymbol{B}_{eb}\right)\boldsymbol{u}_b = \sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \mathbb{C}_e \boldsymbol{\epsilon}_e^*, \tag{2.23a}$$

$$\sum_{b=1}^{n}\left(\sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \mathbb{C}_e \boldsymbol{B}_{eb}\right)\boldsymbol{\eta}_b = \boldsymbol{f}_a - \sum_{e=1}^{m} w_e \boldsymbol{B}_{ea}^T \boldsymbol{\sigma}_e^*. \tag{2.23b}$$

Here we recognize two standard linear-elastic equilibrium problems with identical stiffness matrix corresponding to the reference linear solid defined by W_e and W_e^*, $e = 1, \ldots, m$. The displacement problem (2.23a) is driven by the optimal local strains, whereas the Lagrange multiplier problem (2.23b) is driven by the out-of-balance forces attendant to the optimal local stresses.

2.3.2 Using material symmetries to reduce data sets

Phase-space sampling requirements can be reduced if *a priori* knowledge of material behavior is available. In particular, material symmetry can be effectively exploited for purposes of reducing material sampling requirements. A simple and commonly encountered example of material symmetry is isotropy. For a three-dimensional isotropic material in the linearized kinematics approximation, if $(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e)$ is a material data point, then so are $(\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}_e \boldsymbol{R}_e, \boldsymbol{R}_e{}^T \boldsymbol{\sigma}_e \boldsymbol{R}_e)$ for all rotation matrices $\boldsymbol{R}_e \in SO(3)$, the group of proper orthogonal matrices in three dimensions. Thus, if a point $(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e)$ is in the local data set E_e, then so the entire *orbit* of the point by $SO(3)$.

Under these conditions, local optimality demands

$$F_e(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e) = \min_{(\boldsymbol{\epsilon}'_e, \boldsymbol{\sigma}'_e) \in E_e} \min_{\boldsymbol{R}_e \in SO(3)} \Big(W_e(\boldsymbol{\epsilon}_e - \boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e) + W_e^*(\boldsymbol{\sigma}_e - \boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e) \Big), \tag{2.24}$$

The corresponding optimality condition is

$$\begin{aligned} &\frac{\partial W_e}{\partial \epsilon_{ij}} \frac{\partial}{\partial R_{mn}} (\epsilon'_{kl} R_{ki} R_{lj}) + \frac{\partial W_e^*}{\partial \sigma_{ij}} \frac{\partial}{\partial R_{mn}} (\sigma'_{kl} R_{ki} R_{lj}) - \\ &\frac{\partial}{\partial R_{mn}} (\Lambda_{ij} R_{ki} R_{kj}) = 0, \end{aligned} \tag{2.25}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^T$ is a Lagrange multiplier enforcing the orthogonality of $\boldsymbol{R}_e$. Evaluating the deriva-

tives, we obtain, in matrix form,

$$\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e \left(\frac{\partial W_e}{\partial \boldsymbol{\epsilon}_e} \right) + \boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e \left(\frac{\partial W_e^*}{\partial \boldsymbol{\sigma}_e} \right) = \boldsymbol{\Lambda}. \tag{2.26}$$

Transposing both sides and using tensor symmetry we obtain

$$\left(\frac{\partial W_e}{\partial \boldsymbol{\epsilon}_e} \right) \boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e + \left(\frac{\partial W_e^*}{\partial \boldsymbol{\sigma}_e} \right) \boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e = \boldsymbol{\Lambda}, \tag{2.27}$$

whence it follows that

$$\begin{aligned} &(\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e) \left(\frac{\partial W_e}{\partial \boldsymbol{\epsilon}_e} \right) + (\boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e) \left(\frac{\partial W_e^*}{\partial \boldsymbol{\sigma}_e} \right) = \\ &\left(\frac{\partial W_e}{\partial \boldsymbol{\epsilon}_e} \right) (\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e) + \left(\frac{\partial W_e^*}{\partial \boldsymbol{\sigma}_e} \right) (\boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e). \end{aligned} \tag{2.28}$$

These equations are now to be solved for the local optimal principal directions $\{\boldsymbol{R}_e,\ e = 1, \ldots, m\}$, e. g., by recourse to a Newton-Raphson iteration based on a convenient parametrization of $SO(3)$.

A simple situation arises when the local state $(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e\}$ is itself isotropic, i. e., $\boldsymbol{\epsilon}_e$ and $\boldsymbol{\sigma}_e\}$ have the same principal directions, and W_e and W_e^* are chosen to be isotropic. In this case, the optimality condition (2.28) is satisfied if $\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e$ and DW_e and $\boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e$ and DW_e^* *commute*, which in turn holds if and only if $\boldsymbol{R}_e{}^T \boldsymbol{\epsilon}'_e \boldsymbol{R}_e$ and $\boldsymbol{\epsilon}_e$ and $\boldsymbol{R}_e{}^T \boldsymbol{\sigma}'_e \boldsymbol{R}_e$ and $\boldsymbol{\sigma}_e$ have the same eigenvectors. Introducing the representations

$$\begin{aligned} \boldsymbol{\epsilon}_e &= \boldsymbol{Q}_e{}^T \boldsymbol{e}_e \boldsymbol{Q}_e, \qquad & \boldsymbol{\sigma}_e &= \boldsymbol{Q}_e{}^T \boldsymbol{s}_e \boldsymbol{Q}_e, \\ \boldsymbol{\epsilon}'_e &= \boldsymbol{Q}'_e{}^T \boldsymbol{e}'_e \boldsymbol{Q}'_e, \qquad & \boldsymbol{\sigma}'_e &= \boldsymbol{Q}'_e{}^T \boldsymbol{s}'_e \boldsymbol{Q}'_e, \end{aligned} \tag{2.29}$$

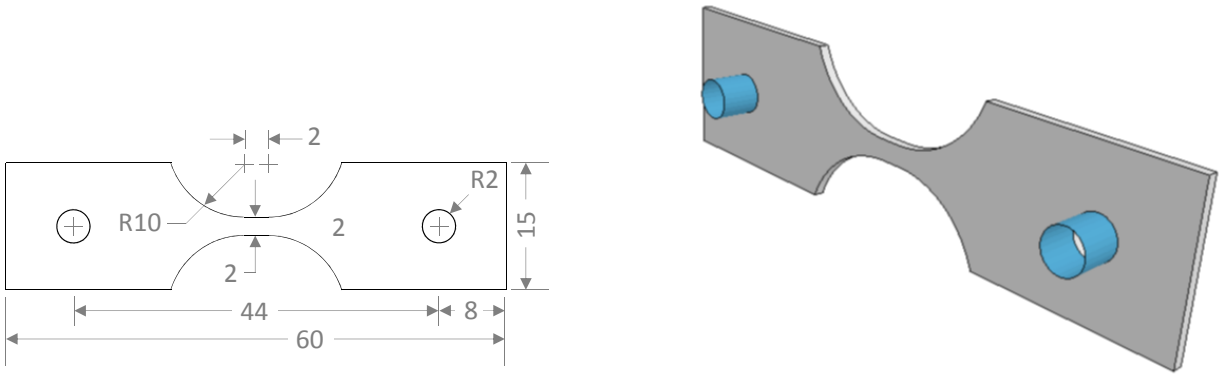
with $\boldsymbol{Q}_e, \boldsymbol{Q}'_e \in SO(3)$ and $\boldsymbol{e}_e$, $\boldsymbol{s}_e$, $\boldsymbol{e}'_e$, $\boldsymbol{s}'_e$ diagonal, local optimality then requires

$$\boldsymbol{R}_e = \boldsymbol{Q}'_e \boldsymbol{Q}_e{}^{-1}, \tag{2.30}$$

which determines explicitly the optimal data point in the $SO(3)$-orbit of $(\boldsymbol{e}'_e, \boldsymbol{s}'_e)$.

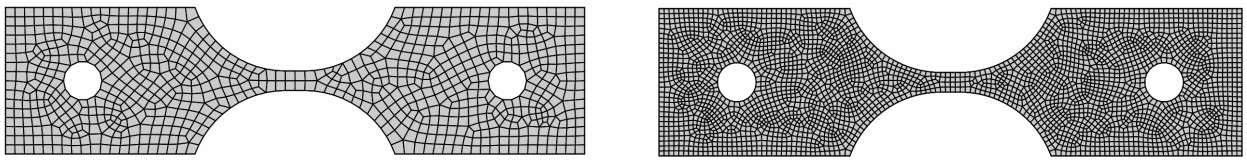
In general, since the local states $(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e\}$ follow from independent Euler-Lagrange equations, eqs. (2.20), they need not be exactly isotropic and the general optimality equations (2.28) need to be solved in order to determine the optimal principal directions of the local data points.

2.3.3 Numerical analysis of convergence



(a) (b)

Figure 2.10: a) Sketch of the simulation set-up of a thin tensile specimen loaded in tension [36]. The thickness of the sample is 1 mm for the three dimensional model. b) Isometric view of the simulation set-up in 3D consisting of two rigid pins and the tensile specimen.



(a) (b)

Figure 2.11: a) Coarse mesh with 811 element and an average element edge length $h \approx 1$mm; b) Fine mesh with 6428 elements and an average element edge length $h = 0.5$mm.

Similarly to the case of truss analysis considered earlier, we revisit the question of convergence of the linear-elasticity data-driven solver with respect to the data set. We specifically consider the problem of the thin tensile specimen shown in Fig. 2.10, cf. [36]. The specimen is loaded by two rigid pins and contains a short gauge section undergoing ostensibly homogeneous deformation. By contrast, the regions surrounding the pin-loaded holes undergo complex heterogeneous deformations. Two finite element discretizations are used in order to ascertain the influence of mesh resolution. The coarse mesh, Fig. 2.11a, consists of 811 elements and one element across the thickness, whereas the fine mesh, Fig. 2.11b, consists of 3,214 elements in two-dimensions and 6,428 elements in three-dimensions, respectively. These discretizations correspond to average element sizes of $h = 1$mm for the coarse mesh and $h = 0.5$mm for the fine mesh. The mesh consists of eight-node hexahedral

elements containing eight Gauss quadrature points each.

Sampling requirements are reduced by virtue of the plane-stress conditions of the problem under consideration. Specifically, only a neighborhood of the subspace $\sigma_{13} = \sigma_{23} = \sigma_{33} = 0$ in stress space needs to be covered by the data. We accomplish this requirement by sampling an appropriate region of the $(\sigma_{11}, \sigma_{22}, \tau_{12})$ stress plane on a uniform cubic grid. The corresponding strains $(\epsilon_{11}, \epsilon_{22}, \epsilon_{12})$ then obey an isotropic linear-elastic law. A reference isotropic linear-elastic solid of the type (2.16), unrelated to the actual material behavior sampled by the material data, is used in the data-driven calculations. These reductions effectively limit the material data set to a three-dimensional space.

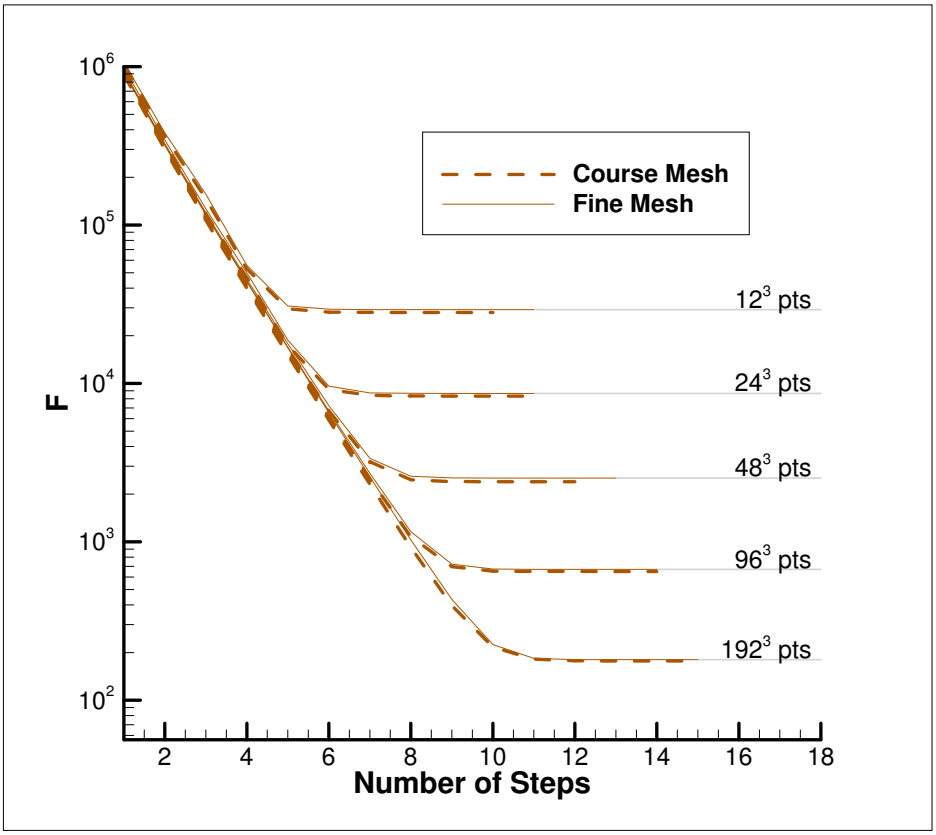


Figure 2.12: Linear-elastic tensile specimen. Convergence of the local material-data assignment iteration. Functional F decays through increasing data resolution in a three dimensional sampling of the plane stress space for both mesh resolutions.

As in the case of the data-driven truss problem, we focus on the questions of convergence for a data-driven linear-elastic solver with respect to local data assignment, or step-wise convergence, and with respect to the data set. In Fig. 2.12, the global functional F is again shown to decay through

iteration for both mesh resolutions on increasingly large data sets. The number of iterations to convergence increases with the material-data sample size but, remarkably, remains modest in all cases.

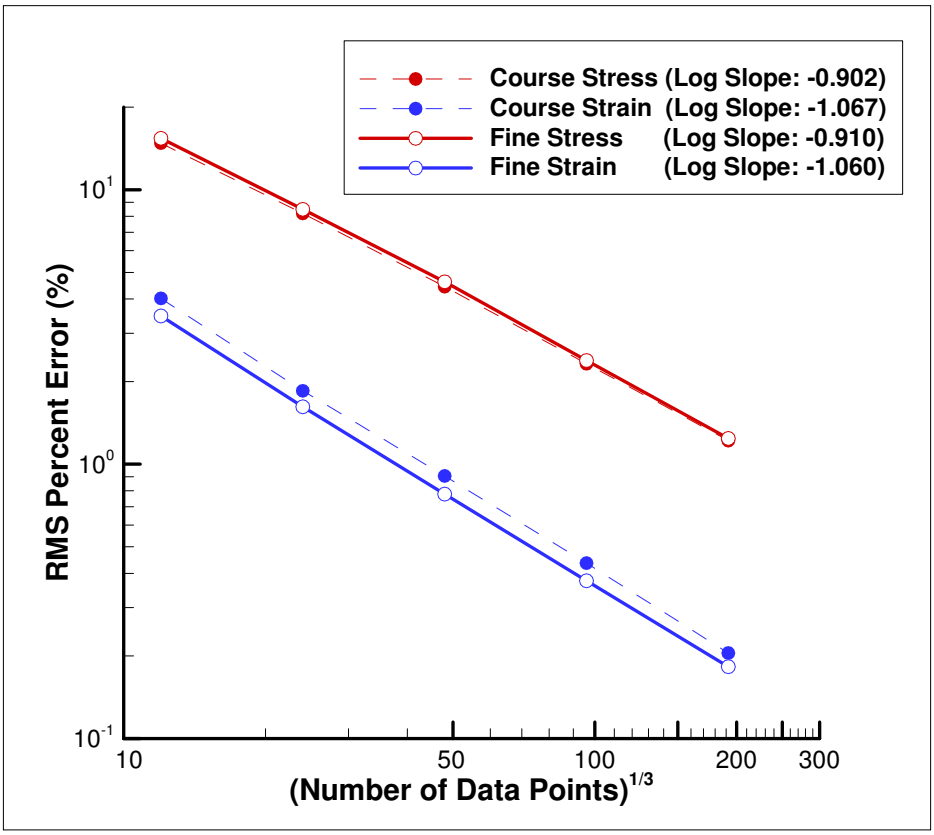


Figure 2.13: Linear-elastic tensile specimen. Convergence with respect to sample size. RMS errors decay linearly in data resolution for both stresses ($\boldsymbol{\sigma}$) and strains ($\boldsymbol{\epsilon}$).

To calculate percent error with respect to the reference solution we re-define the RMS error metric as

$$\sigma_{(\%RMS)} = \left(\frac{\sum_{e=1}^{m} w_e W^* \boldsymbol{\sigma}_e - \boldsymbol{\sigma}_e^{\text{ref}})}{\sum_{e=1}^{m} w_e W^* (\boldsymbol{\sigma}_e^{\text{ref}})} \right)^{1/2}, \tag{2.31a}$$

$$\epsilon_{(\%RMS)} = \left(\frac{\sum_{e=1}^{m} w_e W (\boldsymbol{\epsilon}_e - \boldsymbol{\epsilon}_e^{\text{ref}})}{\sum_{e=1}^{m} w_e W (\boldsymbol{\epsilon}_e^{\text{ref}})} \right)^{1/2}, \tag{2.31b}$$

where W and W^* are the strain and complementary-energy densities as calculated using the reference solution moduli, respectively. Plots of these errors against the cubic root of the number of data points are shown in Fig. 2.13. The plots are indicative of ostensibly linear convergence, in

keeping with the analytical estimate of Corollary 1.

2.4 Mathematical analysis of convergence

We proceed to abstract from the preceding examples a general class of data-driven problems and to establish some of their fundamental properties by way of analysis. We consider systems whose state is characterized by points points in a certain phase space Z. For instance, in the case of linear elasticity, the system of interest is an elastic solid occupying a certain domain Ω and the state of the system is defined by the pair $(\boldsymbol{\epsilon}, \boldsymbol{\sigma})$, where $\boldsymbol{\epsilon}$ is the strain field and $\boldsymbol{\sigma}$ is the stress field, both defined over Ω. In this case, the phase space Z of the elastic solid is an appropriate space of pairs $(\boldsymbol{\epsilon}, \boldsymbol{\sigma})$ of strain and stress fields over Ω.

We particularly wish to characterize the states of the system that are in a constraint set C of states satisfying essential constraints and conservation laws. For instance, in the running example of linear elasticity we may wish to determine states $(\boldsymbol{\epsilon}, \boldsymbol{\sigma}) \in Z$ satisfying compatibility, i. e., such that

$$\boldsymbol{\epsilon}(\boldsymbol{x}) = {}^1\!/_2\big(\nabla \boldsymbol{u}(\boldsymbol{x}) + \nabla \boldsymbol{u}^T(\boldsymbol{x})\big), \qquad \boldsymbol{x} \in \Omega, \tag{2.32a}$$

$$\boldsymbol{u}(\boldsymbol{x}) = \bar{\boldsymbol{u}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega_D, \tag{2.32b}$$

for some displacement field $\boldsymbol{u}$ over Ω and prescribed displacements $\bar{\boldsymbol{u}}$ over the Dirichlet boundary $\partial\Omega_D$, and satisfying equilibrium, i. e., such that

$$\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}) + \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}, \qquad \boldsymbol{x} \in \Omega, \tag{2.33a}$$

$$\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{n}(\boldsymbol{x}) = \bar{\boldsymbol{t}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega_N, \tag{2.33b}$$

for some applied body force field $\boldsymbol{f}$ over Ω and tractions $\bar{\boldsymbol{t}}$ over the Neumann boundary $\partial\Omega_N$, with unit outer normal $\boldsymbol{n}$.

Classically, the problem is closed by putting forth a material law restricting the set of admissible states to a graph E in Z. For instance, in linearized elasticity the material law may take the form a nonlinear Hooke's law

$$\boldsymbol{\sigma}(\boldsymbol{x}) = DW\big(\boldsymbol{\epsilon}(\boldsymbol{x})\big), \qquad \boldsymbol{x} \in \Omega, \tag{2.34}$$

where W is the strain-energy density of the material. The set E then consists of the set of strain and stress fields satisfying the material law at all material points in Ω. The classical solution set is then the intersection $E \cap C$, consisting of states of the system satisfying the essential constraints, the conservation laws and the material law simultaneously. In the case of linear elasticity, the classical solutions would consist of compatible strain fields and equilibrium stress fields satisfying the material law at all material points. In general, the cardinality of the solution set $E \cap C$ depends on the transversality of C with respect to E, depending on which, the solution set may be empty or non-empty, in which latter case the solution set may consist of a single point, corresponding to a unique classical solution, or multiple points.

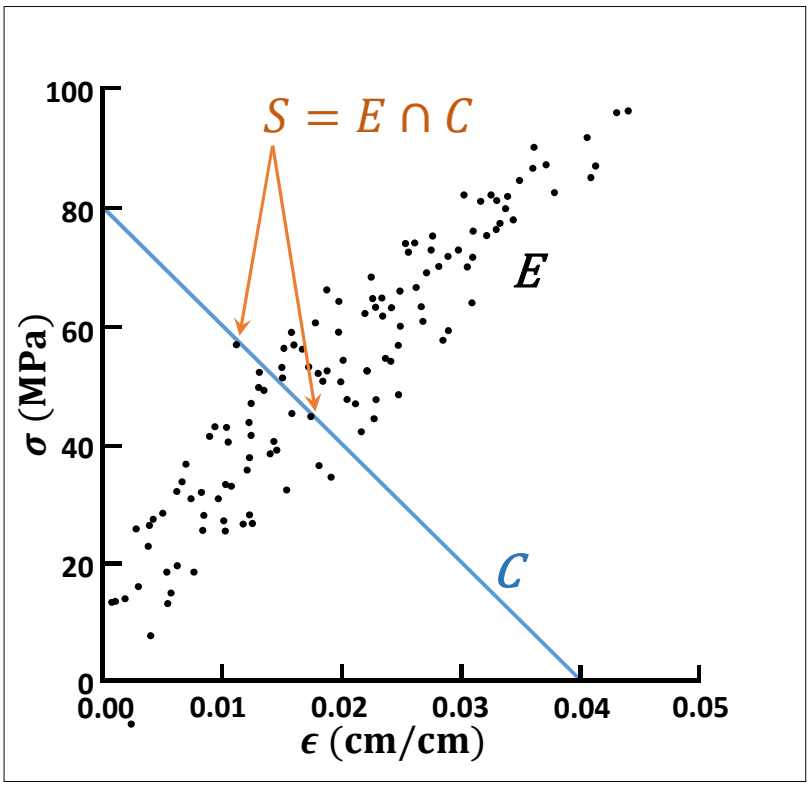


Figure 2.14: Schematic of a local material set E consisting of a finite number of states obtained, e. g., from experimental testing. Also shown is a possible constraint set C and near intersections between E and C.

In contrast to the classical problem just formulated, here we suppose that the material response is not known exactly and, instead, it may be imperfectly characterized by a set, also denoted E, consisting locally at every material point, e. g., of a finite collection of states obtained by means

of experimental testing, cf. Fig. 2.14. Under such conditions, $E \cap C$ is likely to be empty even in cases when solutions could be reasonably expected to exist. It is, therefore, necessary to replace the overly-rigid problem of determining $E \cap C$ by a suitable *penalization* thereof. To this end, we begin by introducing a norm $|\cdot|$ in phase space. For instance, for truss structures we may choose

$$|x|^2 = \sum_{e=1}^{m} w_e \left(C_e \varepsilon_e^2 + \frac{\sigma_e^2}{C_e} \right), \tag{2.35}$$

with $x \equiv (\epsilon_e, \sigma_e)_{e=1}^m$ denoting a generic point in phase space Z. For discretized linear-elastic solids we may choose

$$|x|^2 = \sum_{e=1}^{m} w_e \left(\mathbb{C}_e \boldsymbol{\epsilon}_e \cdot \boldsymbol{\epsilon}_e + \mathbb{C}_e^{-1} \boldsymbol{\sigma}_e \cdot \boldsymbol{\sigma}_e \right), \tag{2.36}$$

with e labeling the integration points in the discretization and $x \equiv (\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e)_{e=1}^m$ denoting a generic point in phase space Z. Finally, for continuum linear elasticity we may choose

$$|x|^2 = \int_\Omega \left(\mathbb{C}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{x}) \cdot \boldsymbol{\epsilon}(\boldsymbol{x}) + \mathbb{C}^{-1}(\boldsymbol{x}) \boldsymbol{\sigma}(\boldsymbol{x}) \cdot \boldsymbol{\sigma}(\boldsymbol{x}) \right) \, dx. \tag{2.37}$$

with $x \equiv \{ \big(\boldsymbol{\epsilon}(\boldsymbol{x}), \boldsymbol{\sigma}(\boldsymbol{x}) \big), \ \boldsymbol{x} \in \Omega \}$ denoting a generic point in phase space Z.

Based on this metrization of phase space, we define the *data-drive problem* as the double minimum problem

$$\min_{y \in E} \min_{x \in C} |y - x| = \min_{y \in E} \operatorname{dist}(y, C), \tag{2.38}$$

or, equivalently,

$$\min_{x \in C} \min_{y \in E} |x - y| = \min_{x \in C} \operatorname{dist}(x, E). \tag{2.39}$$

Thus, the aim of the data-driven problem, as expressed in (2.38), is to *find the point in the material-data set that is closest to satisfying the essential constraints and conservation laws*, or, as expressed in (2.39), to *find the point in the constraint set that is closest to the material-data set*. In the particular example of linear elasticity, the aim of the data-driven problem, as expressed in (2.38), is to find the point in the material-data set that is closest to being compatible and in equilibrium, or, as expressed in (2.39), to find the compatible equilibrium point that is closest to the material-data set.

We note that the data-driven problems considered in Sections 2.2 and Sections 2.3 are indeed

examples of (2.38) and (2.39) with norms (2.35) and (2.36), respectively.

2.4.1 Finite-dimensional case: Convergence with respect to sample size

We begin by considering systems whose local states take values in a finite-dimensional phase space Z. The global state of the system is then characterized by a point $x \in Z$, where m is the number of material points of the system. The essential constraints and conservation laws pertaining to the system have the effect of constraining its global state to lie on a subset C of Z. For instance, for linear-elastic trusses such as considered in the preceding section, the local phase space Z_e of bar e is the space of pairs $(\epsilon, \sigma) \in \mathbb{R}^2$, where ϵ is axial strain of a bar and σ the corresponding axial stress. The global phase space of the entire truss is then $Z = Z_1 \times \cdots \times Z_m$, where m is the number of bars in the truss. In addition, the constraint set C is the affine space of compatible and equilibrated states of stress and strain in the truss.

The data-driven problem (2.38) is now formulated by specifying a set E of possible material states in Z. For instance, in the case of a truss a local material set E_e of the form shown in Fig. 2.14 may be supplying for every bar e of the truss and the global material set is then $E = E_1 \times \cdots \times E_m$. We note that, if E is compact, i. e., consisting of a finite collection of points, then the corresponding data-driven problem has solutions by the Weierstrass extreme-value theorem.

We proceed to consider the question of convergence with respect to the data set. Specifically, we suppose that a sequence (E_k) of data sets is supplied that approximates increasingly closely a limiting data set E. The particular case in which E is a graph concerns convergence of data-driven solutions to classical solutions. For instance, the approximations (E_k) may be the result of an increasing number of experimental tests sampling the behavior of a material characterized by a—possibly unknown—stress-strain curve E. The sequence of approximate material data sets (E_k) generates in turn a sequence of approximate problems

$$\min_{x_k \in C} \min_{y_k \in E_k} |x_k - y_k|, \tag{2.40}$$

and attendant approximate solutions (x_k). We wish to ascertain conditions under which (x_k) converges to solutions of the E-problem.

Conditions ensuring such convergence at a well-defined convergence rate are given in the following proposition. Henceforth, we denote by $\operatorname{dist}(x, E)$ the distance from a point $x \in Z$ to a subset

$E \subset Z$, i. e.,

$$\operatorname{dist}(x, E) = \inf_{y \in E} |x - y|, \tag{2.41}$$

and by $P_Y x$ the projection of $x \in Z$ onto a subspace Y of Z, i. e.,

$$|x - P_Y x| = \operatorname{dist}(x, Y). \tag{2.42}$$

Proposition 1. *Let (E_k) be a sequence of finite subsets of Z, E a subset of Z and C a subspace of Z. Let x be an isolated point of $E \cap C$ and let $x_k, y_k \in Z$ be such that*

$$(x_k, y_k) \in \operatorname{argmin}\{|x - y|,\ x \in C,\ y \in E_k\}. \tag{2.43}$$

Suppose that:

i) There is a sequence $\rho_k \downarrow 0$ such that

$$\operatorname{dist}(z, E_k) \leq \rho_k, \qquad \forall z \in E. \tag{2.44}$$

ii) There is a sequence $t_k \downarrow 0$ such that

$$\operatorname{dist}(z_k, E) \leq t_k, \qquad \forall z_k \in E_k. \tag{2.45}$$

iii) (Transversality) There is a constant $0 \leq \lambda < 1$ and a neighborhood U of x such that

$$|P_C z - x| \leq \lambda |z - x|, \tag{2.46}$$

for all $z \in E \cap U$.

Then,

$$|x_k - x| \leq \frac{t_k + \lambda(t_k + \rho_k)}{1 - \lambda}, \tag{2.47}$$

and, therefore, $\lim_{k \to \infty} |x_k - x| = 0$.

Proof. By assumption (i), we can find $z_k \in E_k$ such that

$$|z_k - x| \leq \rho_k. \tag{2.48}$$

By optimality,

$$\operatorname{dist}(y_k, C) \leq \operatorname{dist}(z_k, C). \tag{2.49}$$

Then, we have

$$|x_k - y_k| = \operatorname{dist}(y_k, C) \leq \operatorname{dist}(z_k, C) \leq |z_k - x| \leq \rho_k. \tag{2.50}$$

By assumption (ii), we can find $z_k \in E$ such that

$$|y_k - z_k| \leq t_k. \tag{2.51}$$

By the triangular inequality, we have

$$|x_k - x| \leq |x_k - P_C z_k| + |P_C z_k - x|. \tag{2.52}$$

By the contractivity of projections,

$$|x_k - P_C z_k| = |P_C y_k - P_C z_k| = |P_C(y_k - z_k)| \leq |y_k - z_k| \leq t_k. \tag{2.53}$$

In addition, by transversality, we have

$$|P_C z_k - x| \leq \lambda |z_k - x|, \tag{2.54}$$

with $0 \leq \lambda < 1$. Triangulating again,

$$|z_k - x| \leq |z_k - y_k| + |y_k - x_k| + |x_k - x|. \tag{2.55}$$

Collecting all the preceding estimates, we obtain

$$|x_k - x| \leq t_k + \lambda(t_k + \rho_k + |x_k - x|), \tag{2.56}$$

whence (2.47) follows. □

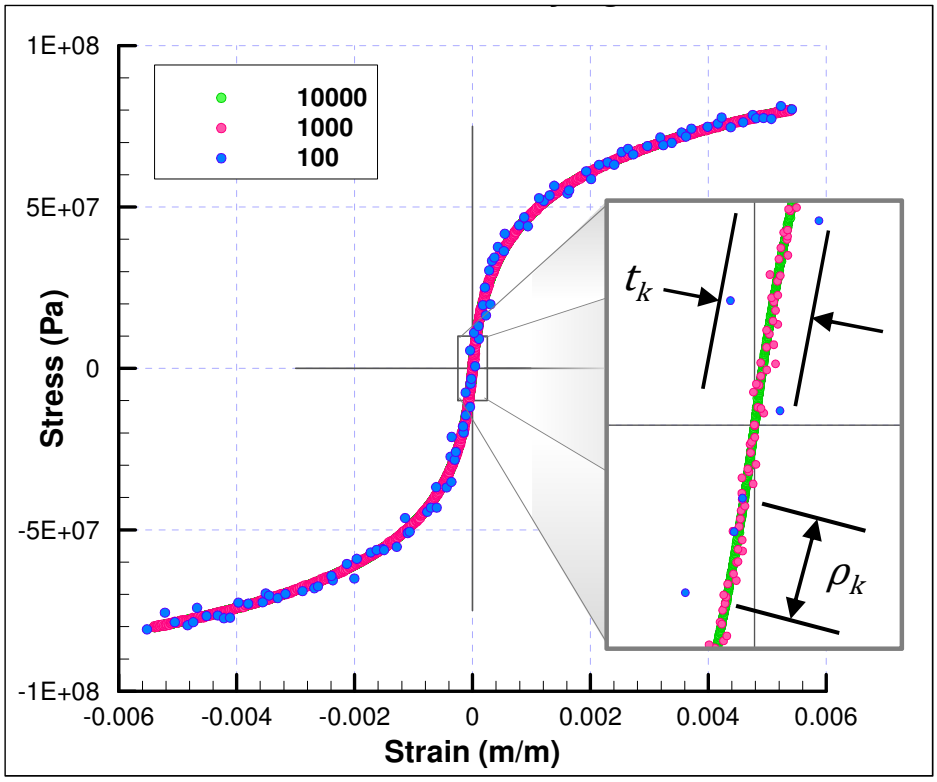


Figure 2.15: Schematic of convergent sequence of material-data sets. The parameter t_k controls the spread of the material-data sets away from the limiting data set and the parameter ρ_k controls the density of material-data point.

The preceding theorem presumes that the classical problem has isolated solutions and that the limiting data set E and the constraint subspace C are transversal at isolated intersections, which are identified with the classical solutions. In particular, E is Lipschitz continuous in a neighborhood of the classical solutions and E is not contained in C in that neighborhood. In the particular case in which the limiting material response is linear, then these conditions reduce to the requirement that the displacement stiffness matrix of the system be non-singular. Assumptions (i) and (ii) set how the sequence of data sets E_k approximate E. Thus, (i) ensures that there are approximate data points increasingly and uniformly closer to any point of E, whereas (ii) ensures that there are no outliers in the approximate data sets such as could spoil the approximate solutions, cf. Fig. 2.15. In particular, (ii) requires E_k to be contained uniformly within t_k-neighborhood of E.

Precise convergence rates with respect to, e. g., the number of sampling points $N_k = |E_k|$, are derived from the preceding theorem if the sequences ρ_k and t_k are related to N_k. In particular, we have the following,

Corollary 1. *Assume that there are constants $C_1 > 0$, $C_2 > 0$ and $\alpha > 0$ such that $\rho_k \leq C_1 N_k^{-\alpha}$ and $t_k \leq C_2 N_k^{-\alpha}$. Then*

$$|x_k - x| \leq \frac{C_2 + \lambda(C_1 + C_2)}{1 - \lambda} N_k^{-\alpha}. \tag{2.57}$$

The numerical convergence rates of Sections 2.2 and 2.3 indeed conform to this estimate. Thus, for the case of truss structures with noise-free data, ρ_k and t_k scale as N_k^{-1}, resulting in a linear convergence rate $\alpha = 1$, cf. Fig. 2.6, whereas for noisy data ρ_k and t_k scale as $N_k^{-1/2}$, resulting in a linear convergence rate $\alpha = 1/2$, cf. Fig. 2.8. For the case of plane-stress linear elasticity with noise-free data, ρ_k and t_k scale as N_k^{-3}, resulting in a linear convergence rate $\alpha = 3$, cf. Fig. 2.13.

2.4.2 Infinite-dimensional case: Convergence with respect to mesh size

The linear-elastic case considered in Section 2.3 differs from the truss case of Section 2.2 in that it obtained by discretization of an infinite-dimensional problem. The question then naturally arises of convergence of the data-driven problem with respect to the mesh size. Consider, for simplicity, a sequence of discretizations of the domain into constant strain triangles of size h_k. Let $x \equiv (\boldsymbol{\epsilon}, \boldsymbol{\sigma})$ be the classical solution and $x_{h_k} \equiv (\boldsymbol{\epsilon}_{h_k}, \boldsymbol{\sigma}_{h_k})$ the corresponding sequence of finite-element solutions. Simultaneously consider a sequence (E_k) of local material-data sets characterizing the satisfying conditions (i) and (ii) of Prop. 1 for some sequences $\rho_k \downarrow 0$ and $t_k \downarrow 0$. Additionally suppose that the sequence of discretizations is regular in the sense that

$$|x_{h_k} - x| \leq C h_k \tag{2.58}$$

for some constant $C > 0$ and that the transversality constants λ_k of the sequence of finite-element models does not degenerate, i. e.,

$$0 \leq \lambda_k \leq \lambda, \tag{2.59}$$

for some $\lambda < 1$. Then, by Prop. 1 we have

$$|x_{k,h_k} - x| \leq |x_{k,h_k} - x_{h_k}| + |x_{h_k} - x| \leq \frac{t_k + \lambda(t_k + \rho_k)}{1 - \lambda} + C h_k, \tag{2.60}$$

where $x_{k,h_k} \equiv (\boldsymbol{\epsilon}_{k,h_k}, \boldsymbol{\sigma}_{k,h_k})$ denotes the data-driven solution corresponding to the E_k material-data set and the h_k discretization. It thus follows that if ρ_k and t_k are controlled by the mesh size, i. e.,

there is a constant $C > 0$ such that

$$\rho_k < Ch_k, \qquad t_k < Ch_k, \tag{2.61}$$

then

$$|x_{k,h_k} - x| \leq Ch_k, \tag{2.62}$$

for some constant $C > 0$ not renamed. We thus conclude that the data-driven paradigm is robust with respect to spatial discretization, in the sense that it preserves convergence provided that the fidelity of the data set increases appropriately with increasing mesh resolution.

2.5 Summary and concluding remarks

We have formulated a new computing paradigm, which we refer to as *data-driven computing*, consisting of formulating calculations *directly* from experimental material data and pertinent essential constraints and conservation laws, thus bypassing the empirical material modeling step of conventional computing altogether. The data-driven solver specifically seeks to assign to each material point of the computational model the closest possible state from a prespecified material-data set, while simultaneously satisfying the conservation laws. Optimality of the local state assignment is understood in terms of a figure of merit that penalizes distance to the data set in phase space. The resulting data-driven problem thus consists of the minimization of a distance function to the data set in phase space subject to constraints set forth by the essential constraints and conservation laws.

We have investigated the performance of the data-driven solver with the aid of two particular examples of application, namely, the static equilibrium of nonlinear three-dimensional trusses and of finite-element discretized linear elastic solids. In these cases, the penalty function in phase space may be regarded as representing a linear-comparison solid with an initial state of strain and stress. The equilibrium constraint can be conveniently enforced by means of Lagrange multipliers. The corresponding stationarity equations correspond to the solution of two linear static equilibrium problems for the comparison solid. We have formulated a local data assignment algorithm by which each member of the truss is pegged to a particular point in the data set. The algorithm terminates when the local state of every member of the truss is in the Voronoi cell of its assigned

data point in phase space. We show, by way of numerical testing, that the data-driven solver has good convergence properties both with respect to the number of data points and with regards to the local data assignment iteration.

The variational structure of the data-driven problem confers robustness to the solver and renders it amenable to analysis. By exploiting this connection, we show that data-driven solutions converge to classical solutions when the data set approximates a limiting constitutive law with increasing fidelity. By virtue of this property, we may regard data-driven problems as a generalization of classical problems in which the material behavior is defined by means of an arbitrary data set in phase space, not necessarily a graph. In particular, classical solutions are recovered precisely when the data set coincides with the graph of a material law.

Whereas the data-driven paradigm has been formulated in the context of computational mechanics and, specifically elastic quasistatic problems, we believe that its range and scope is much larger. Indeed, field theories governed by linear or nonlinear elliptic partial-differential equations should be amenable, upon discretization, to the an analogous treatment. Extensions to dynamical problems are also straightforward. Indeed, dynamics essentially adds inertia forces in the equations of motion that are independent—and do not affect the description—of the material behavior. By contrast, inelastic materials raise the fundamental problem of sampling history-dependent material behavior. Such sampling should provide appropriate coverage of possible processes and evolutions of the system and is thus likely to result in exceedingly large and complex data sets. The use of tools from Data Science and Big Data management may be expected to be particularly beneficial in dealing with such data sets.

We close by pointing out that the traditional computing paradigm has insulated problems from the data on which their solution is based. Removing this barrier creates a powerful new tool in the arsenal of scientific computing. With data-driven computing, data sets can be used directly to provide predictive analysis capability for unmodeled materials. Traceability and inherent measures of data fidelity provide both deeper investigations into data-solution relationships and natural alerts for appropriate model use. Having the ability to tie solution results back to specific data points within a set allows for the creation of a new kind of causality in material analysis. The data-driven paradigm can also ensure the collection of descriptive data sets for prospective uses. Error measures highlight data regions that require additional resolution, as well as point the analyst toward

sensitivities within the solution-source relations. Within the context of conventional computing, these methods can be used to check if a constitutive relation based on a certain data-set is capable of performing a desired simulation *prior to the modeling analysis*. Tying the solution back to the data set also establishes an elegant way of limiting model accuracy to the resolution of the source data. Additionally, material models have specific regimes over which they are developed. However, the models themselves are easily used outside this development range. Especially with regards to empirical ad-hoc curve fits, such overreach can neither be justified nor not easily prevented. By directly using a data set in calculations, attempts to simulate beyond the source data regime are met and penalized by large calculated errors, regardless of how the user receives the data-set. These tangible and intangible benefits add considerable appeal to data-driven solvers beyond their mere usefulness as numerical schemes.

Chapter 3

Entropy Maximizing Data Driven Computing

An earlier version of this work is available as a preprint [57] and has since been submitted for publication [60]. It is presented here with only small modifications.

3.1 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 modelling*. 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 modelling inevitably adds error and uncertainty to the solutions, especially in systems with high-dimensional phase spaces and complex material behavior. This modelling 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.

The current chapter of this thesis is concerned with the further development of Data Driven Computing, tailored to scientific computing and analysis. Data Driven Computing aims to formulate initial-boundary-value problems, and corresponding calculations thereof, directly from material data, thus bypassing the empirical material modelling step of traditional science and engineering altogether. In this manner, material modelling empiricism, error and uncertainty are eliminated entirely and no loss of experimental information is incurred. Here, we extend the work of Chapter 2 to random material data sets with finite probability of *outliers*. We recall that the Data Driven Computing paradigm thus formulated, *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. As was evident, 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. As was shown in Chapter 2, 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 developed in Chapter 2 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 chapter is organized as follows. In Section 3.2, we begin by laying out the connection between Data Science and Scientific Computing that provides the conceptual basis for Data Driven Computing. In Section 3.3, 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 3.4, 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 3.5, 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 3.6.

3.2 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.

3.2.1 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

of the Dirichlet type. The corresponding conjugate variable is the *flux* σ. The flux satisfies the *conservation equation* $\nabla \cdot \sigma = \rho$, where $\nabla\cdot$ is the divergence operator and ρ is a source density, together with natural boundary conditions of the Neumann type. The pair $z = (\epsilon, \sigma)$ describes the local state of the system at a given material point and takes values in the product space $Z = \mathbb{R}^n \times \mathbb{R}^n$, or *phase space*. We note that the phase space, localization, and conservation laws are universal, i. e., material independent. We may thus define a material-independent *constraint set* C to be the set of local states $z = (\epsilon, \sigma)$ consistent with the localization and conservation laws, including corresponding essential and natural boundary conditions.

The localization and conservation laws are closed by appending an appropriate material law. In general, material laws express a relation between fluxes and corresponding *driving forces*. In field theories, the assumption is that local states supply the forces driving the fluxes, leading to material laws of the form $\sigma(\epsilon)$. Often, such material laws are only known imperfectly through a material data set E in phase space Z that collects the totality of our empirical knowledge of the material. Thus, suppose that, in contrast to the classical formulation of initial-boundary-value problems in science and engineering, the material law is imperfectly characterized by a material data point set E. A typical material data set then consists of a finite number of local states, $E = ((\epsilon_i, \sigma_i),\ i = 1, \ldots, N)$. Evidently, for a material data set of this type, the intersection $E \cap C$ is likely to be empty, i. e., there may be no points in the material data set that are compatible with the localization and conservation laws, even in cases when solutions could reasonably be expected to exist. It is, therefore, necessary to replace the overly-rigid characterization of the solution set $S = E \cap C$ by a suitable relaxation thereof.

3.2.2 Distance-minimizing Data Driven schemes

The relaxed formulation of Data Driven Computing developed in Chapter 2 consists of accepting as the best possible solution the point $z_i = (\epsilon_i, \sigma_i)$ in the material data set E that is closest to the constrained set C, i. e., the point that is closest to satisfying the localization and conservation laws. Closeness is understood in terms of some appropriate distance d defined in phase space Z. The corresponding distance from a local state z to the material data set E is then $d(z, E) = \min_{y \in E} d(z, y)$, and the optimal solution is the solution of the minimum problem: $\min_{z \in C} d(z, E)$. Evidently, the data driven problem can also be directly formulated as the double minimization

problem: $\min_{z\in C}\min_{y\in E} d(z,y)$. Inverting the order of minimization, we obtain the equivalent data driven problem: $\min_{y\in E}\min_{z\in C} d(z,y)$, or: $\min_{y\in E} d(y,C)$. This reformulation identifies the data driven solution as the point y in the constraint set C that is closest to the material data set E.

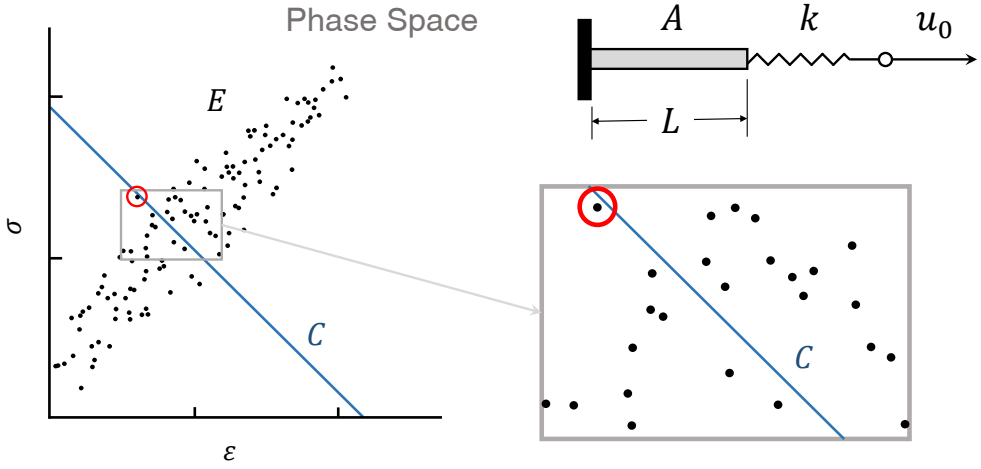


Figure 3.1: Bar loaded by soft device. The data driven solution is the point in the material data set (circled in red) that is closest to the constraint set.

3.2.3 An elementary example

The distance-minimizing Data Driven Computing paradigm just outlined is illustrated in Fig. 3.1 by means of the elementary example of a elastic bar deforming uniformly under the action of a well-calibrated loading device. In this example, phase space Z is the (ϵ,σ)-plane, the material data set is a point set E in phase space and the constraint set is a straight line C of slope and location determined by the stiffness k of the loading device and the applied displacement u_0. In general, the constraint set C and the material data set E may have empty intersection. However, the distance-minimizing Data Driven solution is well defined, though not necessarily uniquely, as the point of the material data set that is closest to the constraint set, circled in red in Fig. 3.1. The fundamental property of the distance-minimizing Data Driven scheme thus defined is that it makes direct use of the material data set in calculations, entirely bypassing the intermediate modelling step of conventional material identification. In the simple example of the bar loaded by a soft device, it is clear that the distance-minimizing Data Driven solution converges to a classical

solution if the data traces a graph in phase space with increasing sampling density, which is a *sanity-check* requirement.

3.2.4 Uniform convergence

The variational structure of distance-minimizing Data Driven problems confers additional robustness to the solvers and renders them amenable to analysis. By exploiting this connection, it was shown in Chapter 2 that distance-minimizing Data Driven solutions converge to classical solutions when the data set approximates a limiting graph in phase space with increasing fidelity, a test case that provides a *sanity check*. Specifically, suppose that the limiting material law is represented by a graph E in phase space, and that a sequence (E_k) of material data sets is such that: i) there is a sequence $\rho_k \downarrow 0$ such that $\text{dist}(z, E_k) \leq \rho_k$, for all $z \in E$, and ii) there is a sequence $t_k \downarrow 0$ such that $\text{dist}(z_k, E) \leq t_k$, for all $z_k \in E_k$. Then, with an additional transversality assumption between the data and the constraint set, it follows from Chapter 2 that the corresponding sequence (z_k) of distance-minimizing Data Driven solutions converges to the solution z of the classical problem defined by the classical material law E. In addition, if the discrete problem is the result of spatial discretization, e. g., by means of the finite element method, then, under the same assumptions, the sequence (u_k) of solutions corresponding to a sequence (E_k) of material data sets converges in norm to the classical solution u of the boundary value problem defined by the classical material law E.

3.3 Probabilistic Data Driven schemes

In practice, material data sets may be random by virtue of inherent stochasticity of the material behavior, experimental scatter inherent to the method of measurement, specimen variability, and other factors. Under these conditions, the material data set may contain outliers, or points far removed from the general clustering of the material data points, with finite probability. If one of these outliers happens to be close to the constraint set, it may unduly dominate the distance-minimizing Data Driven solution described in the foregoing. Thus, such solvers, while well-behaved in applications with material data sets with uniformly bounded scatter, may not be sufficiently robust with respect to persistent outliers in other applications. This limitation of distance-minimizing Data Driven solvers points to the need to investigate problems with random material data sets from a probabilistic perspective, with a view to ranking the data points by relevance and importance and

understanding the probability distribution of outcomes of interest.

3.3.1 Data clustering

Distance-minimizing Data Driven solvers are sensitive to outliers because, for any given test solution z in phase space, they accord overwhelming influence to the nearest point in the material data set, regardless of any clustering of the points. Cluster analysis provides a means of mitigating the influence of individual material data points and building notions of data clustering into the Data Driven solver. Cluster analysis can be based on fundamental concepts of Information Theory such as maximum-entropy estimation [52]. Specifically, we wish to quantify how well a point z in phase space is represented by a point z_i in a material data set $E = (z_1, \dots z_n)$. Equivalently, we wish to quantify the relevance of a point z_i in the material data set to a given point z in phase space. We measure the relevance of points z_i in the material data set by means of weights $p_i \in [0, 1]$ with the property

$$\sum_{i=1}^{n} p_i = 1. \tag{3.1}$$

We wish the ranking by relevance of the material data points to be unbiased. It is known from Information Theory that the most unbiased distribution of weights is that which maximizes Shannon's information entropy [89, 90, 91]

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i \tag{3.2}$$

with the extension by continuity: $0 \log 0 = 0$. In addition, we wish to accord points distant from z less weight than nearby points, i. e., we wish the cost function

$$U(z, p) = \sum_{i=1}^{n} p_i d^2(z, z_i) \tag{3.3}$$

to be as small as possible. These competing objectives can be combined in the sense of Pareto optimality. The Pareto optima are solutions of the problem

$$\text{For fixed } z, \text{ minimize:} \qquad \beta U(z, p) - H(p) \tag{3.4}$$

$$\text{subject to:} \qquad p_i \geq 0,\ i = 1, \dots, n; \quad \sum_{i=1}^{N} p_i = 1, \tag{3.5}$$

where $\beta \in (0, +\infty)$ is a Pareto weight. The solution to this problem is given by the Bolzmann distribution

$$p_i(z, \beta) = \frac{1}{Z(z, \beta)} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}, \tag{3.6a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}. \tag{3.6b}$$

The corresponding max-ent Data Driven solver now consists of minimizing the free energy

$$F(z, \beta) = -\frac{1}{\beta} \log Z(z, \beta), \tag{3.7}$$

over the constraint set C, i. e.,

$$z \in \operatorname{argmin}\{F(z', \beta),\ z' \in C\}. \tag{3.8}$$

We note that $\beta^{-1/2}$ represents the width of the Bolzmann distribution (3.6) in phase space. Thus, points in the data set at a distance to z large compared to $\beta^{-1/2}$ have negligible influence over the solution. Conversely, the solution z is dominated by the local cluster of data points in the $\beta^{-1/2}$-neighborhood of z. In particular, outliers, or points outside that neighborhood, have negligible influence over the solution.

For a compact material point set E in a finite-dimensional phase space, the existence of solutions of problem (3.8) is ensured by the Weierstrass extreme value theorem. We also note that the distance-minimizing Data Driven scheme from Chapter 2 is recovered in the limit of $\beta \to +\infty$. By analogy to statistical thermodynamics, max-ent Data Driven Computing may be regarded as a *thermalized* extension of distance-minimizing Data Driven Computing. For finite β, all points in the material data set influence the solution, but their corresponding weights diminish with distance to the solution. In particular, the addition of an outlier that is marginally closer to the constraint set C than a large cluster of material data points does not significantly alter the solution, as desired.

3.4 Numerical implementation

We recall that the max-ent Data Driven problem of interest is to minimize the free energy $F(z)$ (3.7) over the constraint set C. The corresponding optimality condition is

$$\frac{\partial F}{\partial z}(z,\beta) \perp C, \tag{3.9}$$

where $\perp$ denotes orthogonality. Assuming

$$d(z,z') = |z - z'|, \tag{3.10}$$

with $|\cdot|$ the standard norm in $\mathbb{R}^n$, we compute

$$\frac{\partial F}{\partial z}(z,\beta) = \sum_{i=1}^{n} p_i(z,\beta)(z - z_i) = z - \sum_{i=1}^{n} p_i(z,\beta) z_i. \tag{3.11}$$

Inserting this identity into (3.9), we obtain

$$z - \sum_{i=1}^{n} p_i(z,\beta) z_i \perp C, \tag{3.12}$$

which holds if and only if

$$z = P_C \left(\sum_{i=1}^{n} p_i(z,\beta) z_i \right), \tag{3.13}$$

where P_C is the closest-point projection to C. For instance, if $C = \{f(z) = 0\}$ for some constraint function $f(z)$, (3.9) may be expressed as

$$\frac{\partial F}{\partial z}(z,\beta) = \lambda \frac{\partial f}{\partial z}(z), \tag{3.14a}$$

$$f(z) = 0, \tag{3.14b}$$

where λ is a Lagrange multiplier.

The essential difficulty inherent to problem (3.9), or (3.14), is that, in general, the free energy function $F(\cdot,\beta)$ is strongly non-convex, possessing multiple wells centered at the data points in the material data set. Under these conditions, iterative solvers may fail to converge or may return a

local minimizer, instead of the global minimizer of interest.

We overcome these difficulties by recourse to *simulated annealing* [61]. The key observation is that the free energy $F(\cdot, \beta)$ is convex for sufficiently small β. Indeed, a straightforward calculation using (3.10) gives the Hessian matrix as

$$\begin{aligned}\frac{\partial^2 F}{\partial z \partial z}(z) = I - \beta \sum_{i=1}^{n} p_i(z)(z - z_i) \otimes (z - z_i) \\ + \beta \left(\sum_{i=1}^{n} p_i(z)(z - z_i) \right) \otimes \left(\sum_{j=1}^{n} p_j(z)(z - z_j) \right).\end{aligned} \tag{3.15}$$

Evidently, in the limit of $\beta \to 0$ the Hessian reduces to the identity and the free energy is convex. Indeed, it follows from (3.6b) that, for $\beta \to 0$,

$$F(z, \beta) - \frac{1}{\beta} \log \frac{1}{n} \sim \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} d^2(z, z_i). \tag{3.16}$$

The main idea behind simulated annealing is, therefore, to initially set β sufficiently small that $F(\cdot, \beta)$ is convex and subsequently increase it according to some appropriate annealing schedule, with a view to guiding the solver towards the absolute minimizer.

3.4.1 Fixed-point iteration

We begin by noting that, for fixed β, eq. (3.13) conveniently defines the following fixed-point iteration,

$$z^{(k+1)} = P_C \left(\sum_{i=1}^{n} p_i(z^{(k)}, \beta) z_i \right). \tag{3.17}$$

We recall that fixed-point iterations $z \leftarrow f(z)$ converge if the mapping $f(z)$ is contractive. Since P_C is an orthogonal projection, it is contractive if the constraint set C is convex, which we assume henceforth. Under this assumption, the mapping (3.17) is contractive if and only if the mapping

$$z \mapsto \sum_{i=1}^{n} p_i(z, \beta) z_i \equiv g(z, \beta) = z - \frac{\partial F}{\partial z}(z, \beta) \tag{3.18}$$

is contractive.

Conditions ensuring the contractivity of $g(\cdot, \beta)$ are given by the following theorem.

Theorem 1. *Suppose that*

$$\frac{1}{\beta} < \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2, \tag{3.19}$$

where

$$\bar{z} = \sum_{i=1}^{n} p_i(z,\beta) z_i. \tag{3.20}$$

Then, $g(\cdot,\beta)$ is contractive in a neighborhood of z.

Proof. From the definition (3.18) of $g(z,\beta)$, we have, after a trite calculation,

$$\frac{\partial g}{\partial z}(z) = \beta \sum_{i=1}^{n} p_i(z,\beta)(z_i - \bar{z}) \otimes (z_i - \bar{z}). \tag{3.21}$$

Let $u \in Z$, $|u| = 1$. Then,

$$u^T \frac{\partial g}{\partial z}(z) u = \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z_i - \bar{z}) \cdot u\big)^2 \le \beta \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2. \tag{3.22}$$

Therefore, by the implicit function theorem, contractivity in a neighborhood of z follows if

$$\beta \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2 < 1, \tag{3.23}$$

or, equivalently, if (3.19) holds. □

Conditions ensuring global contractivity of the mapping $g(\cdot,\beta)$ are given by the following theorem.

Theorem 2. *Suppose that*

$$\frac{1}{\beta} > \frac{1}{n} \sum_{i=1}^{n} |z - z_i|^2, \tag{3.24}$$

for all $z \in \Omega \subset Z$. Then, $F(\cdot,\beta)$ is convex in Ω.

Proof. Fix $z \in \Omega$ and $\beta > 0$ and let $u \in Z$ be an arbitrary unit vector in phase space. We have

$$u^T \frac{\partial^2 F}{\partial z \partial z} u = 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i) \cdot u\big)^2 + \beta \left(\sum_{i=1}^{n} p_i(z,\beta)(z - z_i) \cdot u \right)^2, \tag{3.25}$$

which gives the lower bound

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i)\cdot u\big)^2. \tag{3.26}$$

Maximizing the bound with respect to u, we have

$$\sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i)\cdot u\big)^2 \leq \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2, \tag{3.27}$$

and, hence,

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2. \tag{3.28}$$

From the max-ent optimality of $p_i(z,\beta)$, we additionally have

$$\frac{\beta}{2}\sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 + \sum_{i=1}^{n} p_i(z,\beta)\log p_i(z,\beta) \leq \frac{\beta}{2}\sum_{i=1}^{n} p_i'|z - z_i|^2 + \sum_{i=1}^{n} p_i'\log p_i', \tag{3.29}$$

for all (p_i') such that

$$p_i' \geq 0, \quad \sum_{i=1}^{n} p_i' = 1. \tag{3.30}$$

Testing with $p_i' = 1/n$, we obtain

$$\frac{\beta}{2}\sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 + \sum_{i=1}^{n} p_i(z,\beta)\log p_i(z,\beta) \leq \frac{\beta}{2}\frac{1}{n}\sum_{i=1}^{n}|z - z_i|^2 + \log\frac{1}{n}. \tag{3.31}$$

But, by Jensen's inequality,

$$\log\frac{1}{n} \leq \sum_{i=1}^{n} p_i(z,\beta)\log p_i(z,\beta), \tag{3.32}$$

which, in conjunction with (3.31), gives

$$\sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 \leq \frac{1}{n}\sum_{i=1}^{n}|z - z_i|^2. \tag{3.33}$$

Inserting this estimate in (3.28) gives

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \frac{\beta}{n}\sum_{i=1}^{n}|z - z_i|^2. \tag{3.34}$$

From this inequality we conclude that

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 0 \tag{3.35}$$

for all unit vectors u in phase space if

$$1 - \frac{\beta}{n} \sum_{i=1}^{n} |z - z_i|^2 \geq 0, \tag{3.36}$$

or, equivalently, if inequality (3.24) is satisfied. □

3.4.2 Simulated annealing

The general idea of simulated annealing is to evolve the reciprocal temperature jointly with the fixed point iteration according to an appropriate annealing schedule, i. e., we modify (3.17) to

$$z^{(k+1)} = P_C \left(\sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)}) z_i \right). \tag{3.37}$$

An effective annealing schedule is obtained by selecting $\beta^{(k+1)}$ so as to ensure local contractivity of the fixed-point mapping. An appeal to Theorem 1 suggests schedules such that

$$\frac{1}{\beta^{(k+1)}} < \sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)}) |z_i - \bar{z}^{(k)}|^2. \tag{3.38}$$

By Theorem 1, this choice ensures local contractivity and, hence, convergence of the fixed-point iteration (3.17). The initial reciprocal temperature $\beta^{(0)}$ may be chosen according to Theorem 2, which ensures that the fixed-point iteration is contractive everywhere.

As already noted, the max-ent Data Driven solution is controlled by its local $\beta^{-1/2}$-neighborhood of points in the data set. Thus, initially the annealing schedule casts a broad net and all points in the data set are allowed to influence the solution. As β grows, that influence is restricted to an increasingly smaller cluster of data points. For large β, the solution is controlled by the points in a certain local neighborhood of the data set determined by the annealing iteration. In particular, the influence of outliers in the data set is eliminated.

3.5 Numerical tests

We test the properties of max-ent Data Driven Computing by means of the simple example of truss structures. Trusses are assemblies of articulated bars that deform in uniaxial tension or compression. Thus, conveniently, in a truss the material behavior of a bar e is characterized by a simple relation between the uniaxial strain ε_e and uniaxial stress σ_e in the bar. We refer to the space of pairs $z_e = (\varepsilon_e, \sigma_e)$ as the *phase space* of bar e. The state $z = (z_e)_{e=1}^m$, where m is the number of bars in the truss, is subject to the compatibility and equilibrium constraints

$$\epsilon_e = B_e u, \tag{3.39a}$$

$$\sum_{e=1}^{m} B_e^T w_e \sigma_e = f, \tag{3.39b}$$

where u is the array of nodal displacements, f is the array of applied nodal forces, the matrices $(B_e)_{e=1}^m$ encode the geometry and connectivity of the truss members and w_e is the volume of member e.

We may metrize the local phase spaces of each member of the truss by means of Euclidean distances derived from the norms

$$|z_e|_e = \left(\mathbb{C}\epsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\right)^{1/2}, \tag{3.40}$$

for some positive constant $\mathbb{C}$. We may then metrize the global state of the truss by means of the global norm

$$|z| = \Big(\sum_{e=1}^{m} w_e |z_e|_e^2\Big)^{1/2} = \left(\sum_{e=1}^{m} w_e \Big(\mathbb{C}\epsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\Big)\right)^{1/2} \tag{3.41}$$

and the associated distance (3.10). For a truss structure, the point in C closest to a given point z^* in phase space follows from the stationarity condition

$$\delta\left\{\sum_{e=1}^{m} w_e \left(\frac{\mathbb{C}}{2}(B_e u - \epsilon_e^*)^2 + \frac{\mathbb{C}^{-1}}{2}(\sigma_e - \sigma_e^*)^2\right) + \Big(f - \sum_{e=1}^{m} w_e B_e^T \sigma_e\Big)^T \lambda\right\} = 0, \tag{3.42}$$

where λ is an array of Lagrange multiplier enforcing the equilibrium constraints. The corresponding

Euler-Lagrange equations are

$$\sum_{e=1}^{m} w_e B_e^T \mathbb{C}(B_e u - \epsilon_e^*) = 0, \tag{3.43a}$$

$$\mathbb{C}^{-1}(\sigma_e - \sigma_e^*) - B_e \lambda = 0, \tag{3.43b}$$

$$\sum_{e=1}^{m} w_e B_e^T \sigma_e = f, \tag{3.43c}$$

or

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) u = \sum_{e=1}^{m} w_e B_e^T \mathbb{C} \epsilon_e^*, \tag{3.44a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) \lambda = f - \sum_{e=1}^{m} w_e B_e^T \sigma_e^*, \tag{3.44b}$$

which define two standard truss equilibrium problems for the linear reference material of modulus $\mathbb{C}$.

We assume that the behavior of each bar e is characterized by a local material data set $E_e = \{z_{i_e} = (\epsilon_{i_e}, \sigma_{i_e}) \in \mathbb{R}^2,\ i_e = 1, \ldots, n_e\}$, where n_e is the number of data points in E_e. For instance, each point in the data set may correspond, e. g., to an experimental measurement. A typical data set is notionally depicted in Fig. 3.1. The global data set is then the Cartesian product

$$E = E_1 \times \cdots \times E_m. \tag{3.45}$$

A typical point in such a data set is most convenient indexed as $z_{i_1 \ldots i_m}$, with $i_e = 1, \ldots, n_e$, $e = 1, \ldots, m$, instead of using a single index as in Section 3.3. The partition function (3.6b) then takes the form

$$Z(z, \beta) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \mathrm{e}^{-(\beta/2) \sum_{e=1}^{m} d^2(z_e, z_{i_e})}, \tag{3.46}$$

where the local distance is given by (3.40). Rearranging terms, (3.46) may be rewritten in the form

$$\begin{aligned} Z(z, \beta) &= \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \left(\prod_{e=1}^{m} \mathrm{e}^{-(\beta/2) d^2(z_e, z_{i_e})}\right) \\ &= \prod_{e=1}^{m} \left(\sum_{i_e=1}^{n_e} \mathrm{e}^{-(\beta/2) d^2(z_e, z_{i_e})}\right) \equiv \prod_{e=1}^{m} Z_e(z_e, \beta), \end{aligned} \tag{3.47}$$

and the total free energy evaluates to

$$F(z,\beta) = -\frac{1}{\beta}\log Z(z,\beta) = \sum_{e=1}^{m}\left(-\frac{1}{\beta}\log Z_e(z_e,\beta)\right) \equiv \sum_{e=1}^{m} F_e(z_e,\beta). \tag{3.48}$$

We note that the total free energy is additive with respect to the free energies $F_e(z_e,\beta)$ of the members. Finally, the Bolzmann distribution (3.6) becomes

$$\begin{aligned} p_{i_1,\ldots,i_m}(z,\beta) &= \frac{1}{Z(z,\beta)}\mathrm{e}^{-(\beta/2)\sum_{e=1}^{m} d^2(z_e,z_{i_e})} \\ &= \prod_{e=1}^{m}\left(\frac{1}{Z_e(z_e,\beta)}\mathrm{e}^{-(\beta/2)d^2(z_e,z_{i_e})}\right) \equiv \prod_{e=1}^{m} p_{i_e}(z_e,\beta). \end{aligned} \tag{3.49}$$

As expected, the local memberwise probability distributions are independent. We also note that the system is assumed to be in thermal equilibrium, i. e., the members are all assumed to be at the same temperature.

In the case of independent local material data sets, eq. (3.45), the bound (3.19) specializes to

$$\begin{aligned} \frac{1}{\beta} &< \sum_{i_1=1}^{n_1}\cdots\sum_{i_m=1}^{n_m} p_{i_1,\ldots,i_m}(z,\beta)\left(\sum_{e=1}^{m} d^2(\bar{z}_e,z_{i_e})\right) \\ &= \sum_{e=1}^{m}\left(\sum_{i_1=1}^{n_1}\cdots\sum_{i_m=1}^{n_m} p_{i_1,\ldots,i_m}(z,\beta)d^2(\bar{z}_e,z_{i_e})\right) \\ &= \sum_{e=1}^{m}\left(\sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta)d^2(\bar{z}_e,z_{i_e})\right). \end{aligned} \tag{3.50}$$

We can exploit this special structure and refine the bound by applying it at the local level, i. e., by requiring

$$\frac{1}{\beta_e} < \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta_e)d^2(\bar{z}_e,z_{i_e}), \tag{3.51}$$

$e = 1,\ldots,m$, where $1/\beta_e$ represent local temperatures. We can further define an annealing schedule by taking (3.51) as the basis for local temperature updates

$$\frac{1}{\beta_e^{(k+1)}} = \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta^{(k)})d^2(\bar{z}_e^{(k)},z_{i_e}), \tag{3.52}$$

with thermal equilibrium subsequently restored by setting the global temperature to

$$\frac{1}{\beta^{(k+1)}} = \sum_{e=1}^{m} \frac{w_e^{(k+1)}}{\beta_e^{(k+1)}}, \tag{3.53}$$

with appropriate weights $w_e^{(k+1)}$. In calculations, we specifically choose

$$w_e^{(k+1)} = \frac{\mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}}{\sum_{e=1}^{m} \mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}} = \frac{Z_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}{\sum_{e=1}^{m} Z_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}. \tag{3.54}$$

Finally, the initial estimate (3.24) corresponds to setting

$$p_{i_e}(z_e, \beta) = \frac{1}{n_e}, \tag{3.55}$$

whereupon (3.38) becomes

$$\frac{1}{\beta^{(0)}} = \sum_{e=1}^{m} \frac{1}{n_e} \left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e}) \right). \tag{3.56}$$

As a further control on the annealing rate we set

$$\beta^{(k+1)} = \lambda \tilde{\beta}^{(k+1)} + (1 - \lambda)\beta^{(k)}, \tag{3.57}$$

where $\tilde{\beta}^{(k+1)}$ is the result of applying the update (3.53) and λ is an adjustable factor.

A complete list of relations specialized for the case of independent local material data sets is given in Algorithm 2.

Alternative strategies for starting and accelerating simulated-annealing iterations are briefly noted in Section 3.6.5, but a detailed investigation of such alternatives is beyond the scope of this thesis (cf., e. g., [66] for a general discussion of simulated-annealing strategies). The iterative solver operates in two distinct phases. The first phase executes the annealing schedule until the values for β become large. Subsequent to this initial phase, the algorithm proceeds by distance minimization, as in Chapter 2, until convergence is achieved.

Algorithm 2 Data Driven solver

Require: Local data sets $E_e = \{z_{i_e},\, i_e = 1, \dots, n_e\}$, B-matrices $\{B_e,\, e = 1, \dots, m\}$, force vector f, parameter λ.

1) Initialize data iteration. Set $k = 0$, compute

$$\bar{z}_e^{(0)} = z_e^{(0)} = \frac{1}{n_e} \sum_{i_e=1}^{n_e} z_{i_e}, \quad \frac{1}{\beta^{(0)}} = \sum_{e=1}^{m} \frac{1}{n_e} \left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e}) \right). \tag{3.58}$$

2) Calculate data associations and precalculate for convexity estimate:

for all $e = 1, \dots, m$ **do**

2.1) Set $c_{i_e}^{(k)} = \exp\big(-\beta^{(k)} d^2(z_e^{(k)}, z_{i_e}) \big),\ i_e = 1, \dots, n_e$.

2.2) Set $Z_e^{(k)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(k)}$.

2.3) Set $p_{i_e}^{(k)} = c_{i_e}^{(k)} / Z_e^{(k)},\ i_e = 1, \dots, n_e$.

2.4) Set $\bar{z}_e^{(k)} = \sum_{i_e=1}^{n_e} p_{i_e}^{(k)} z_{i_e}$.

2.5) Set $D_e^{(k)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(k)} d^2(\bar{z}_e^{(k)}, z_{i_e})$

end for

3) Solve for $u^{(k+1)}$ and $\eta^{(k+1)}$:

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e \right) u^{(k+1)} = \sum_{e=1}^{m} w_e B_e^T \mathbb{C} \bar{\epsilon}_e^{(k)}, \tag{3.59a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e \right) \eta^{(k+1)} = f - \sum_{e=1}^{m} w_e B_e^T \bar{\sigma}_e^{(k)}. \tag{3.59b}$$

4) Progress Schedule:

4.1) Set

$$\tilde{\beta}^{(k+1)} = \left(\frac{\sum_{e=1}^{m} D_e^{(k)}}{\sum_{e=1}^{m} Z_e^{(k)}} \right)^{-1}. \tag{3.60}$$

4.2) Set $\beta^{(k+1)} = (1 - \lambda)\beta^{(k)} + \lambda \tilde{\beta}^{(k+1)}$.

5) Compute local states $z_{e,k}$:

for all $e = 1, \dots, m$ **do**

$$\varepsilon_e^{(k+1)} = B_e u^{(k+1)}, \qquad \sigma_e^{(k+1)} = \bar{\sigma}_e^{(k+1)} + \mathbb{C} B_e \eta^{(k+1)} \tag{3.61}$$

end for

6) Test for convergence and cycle the time or data iteration:

if $\{z_e^{(k+1)} = z_e^{(k)},\ e = 1, \dots, m\}$ **then**

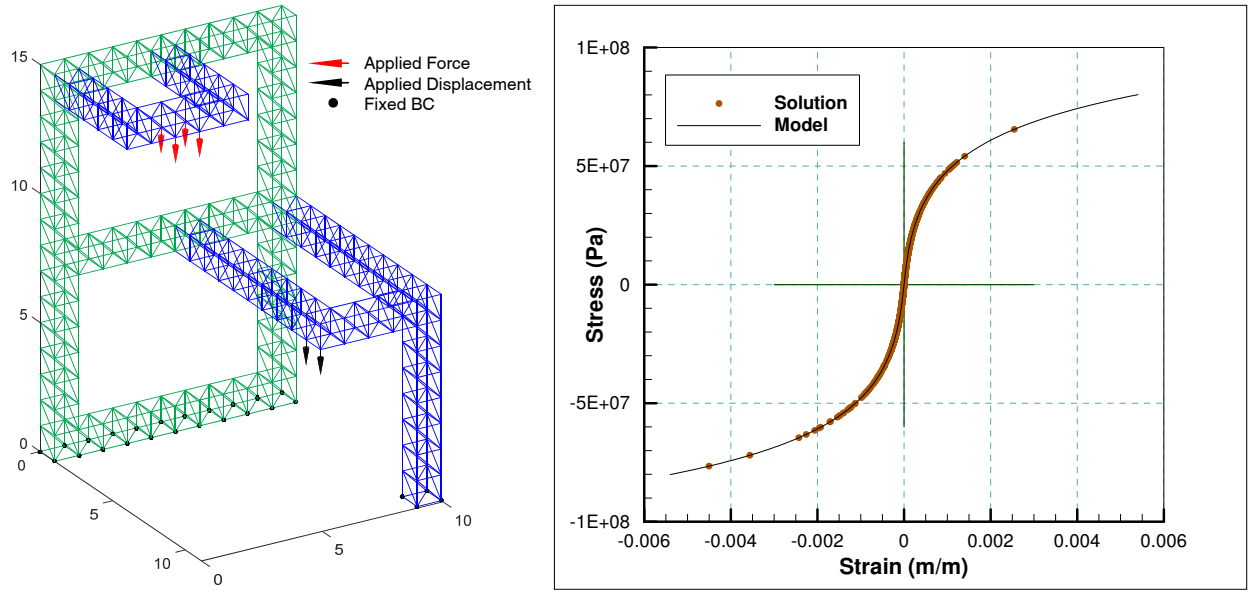
exit

else

$k \leftarrow k + 1$,

goto (2).

end if

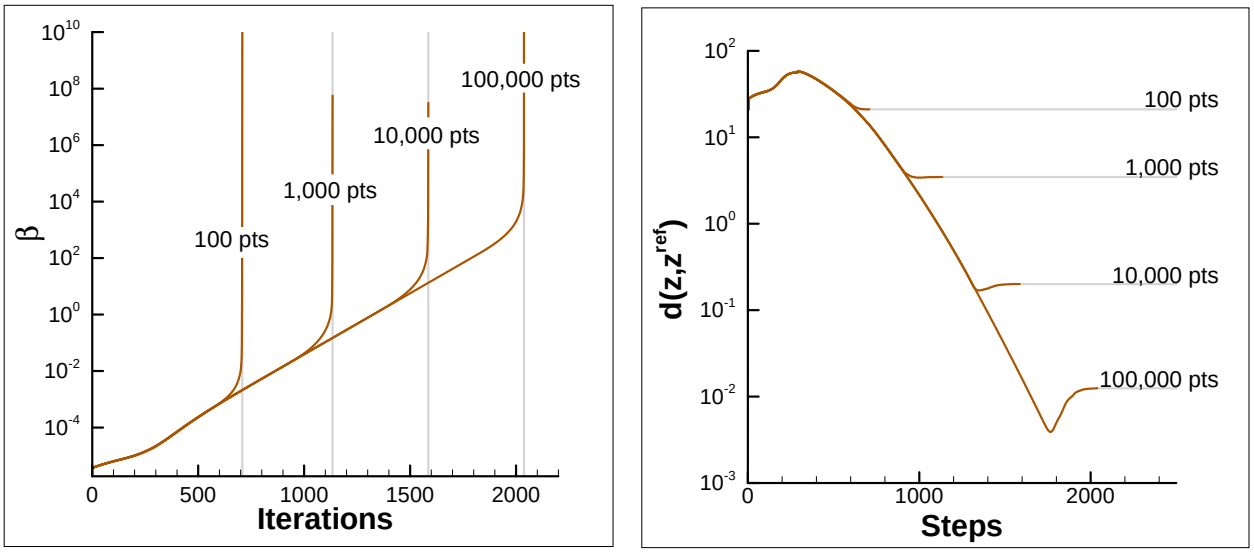


(a) (b)

Figure 3.2: a) Geometry and boundary conditions of truss test case. b) Base material model with reference solution stress-strain points superimposed.

3.5.1 Annealing schedule

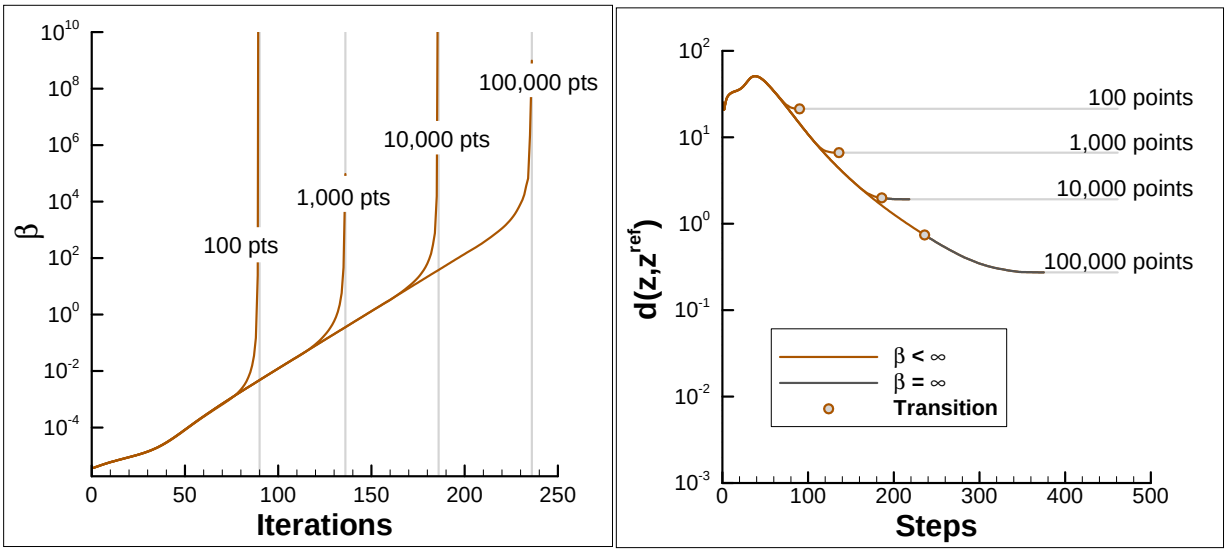
In calculations we consider the specific test case shown in Fig. 3.2a. The truss contains 1,246 members and is supported and loaded as shown in the figure. By way of reference, we consider the nonlinear stress-strain relation shown in 3.2b. A Newton-Raphson solution based on that model is readily obtained. The resulting states of all the members of the truss are shown in Fig. 3.2a superimposed on the stress-strain curve in order to visualize the coverage of phase space entailed by the reference solution.



(a) (b)

Figure 3.3: Truss test case, , $\lambda = 0.01$. a) Evolution of β through the annealing schedule for different data set sizes. b) Convergence of the max-ent Data Driven solution to the reference solution for the base model depicted in Fig. 3.2b.

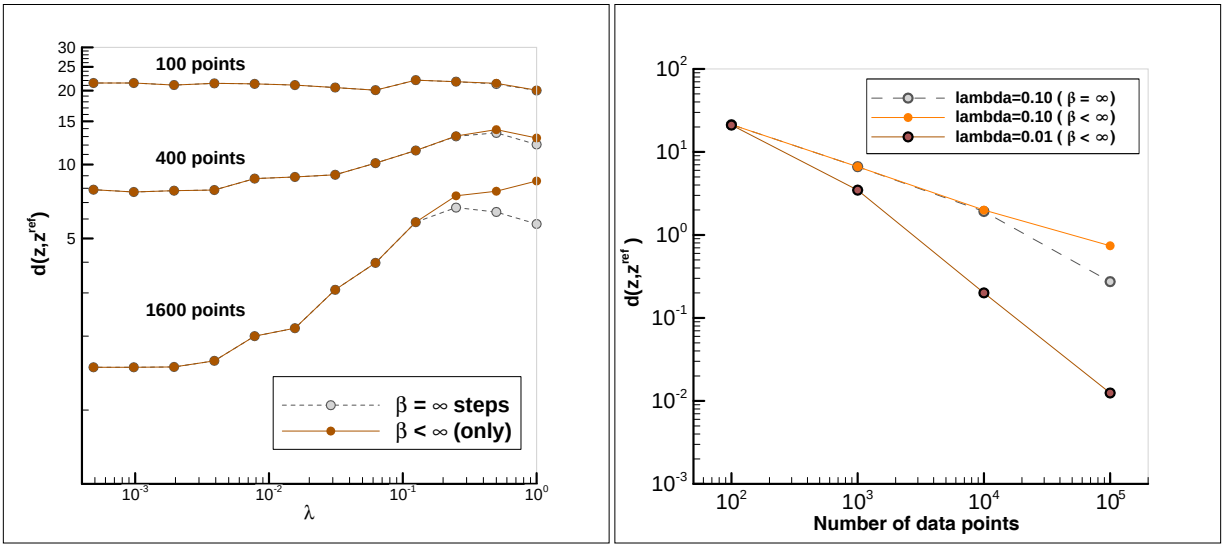
The performance of the solver is shown in Fig. 3.3. In the present test, data sets are generated by spacing the data points evenly over the strain axis and then evaluating the corresponding stress values from the base model depicted in Fig. 3.2b. Fig. 3.3a shows the evolution of β through the annealing schedule for $\lambda = 0.01$. As may be seen from the figure, β grows roughly linearly up to a certain, data set size dependent, number of iterations at which point it diverges rapidly. The stepwise convergence of the simulated annealing iteration is shown in Fig. 3.3b. As the data set grows in size, the number of iterations to convergence grows correspondingly, as the iteration has to explore a larger data set.



(a) (b)

Figure 3.4: Truss test case, $\lambda = 0.1$. a) Evolution of β through the annealing schedule for different data set sizes. b) Convergence of the max-ent Data Driven solution to the reference solution for the base model depicted in Fig. 3.2b.

The influence of the parameter λ on the annealing schedule and the solution is illustrated in Fig. 3.4, which corresponds to $\lambda = 0.1$. In general, a larger value of λ represents a more aggressive, or faster, annealing schedule, whereas a smaller value represents a more conservative, or slower, annealing schedule. A comparison between Figs. 3.3 and 3.4 reveals that, whereas an aggressive annealing schedule indeed speeds up the convergence of the simulated-annealing iteration, it may prematurely freeze the solution around a non-optimal data set cluster, with an attendant loss of accuracy of the solution. Contrariwise, whereas a conservative annealing schedule slows down the convergence of the simulated-annealing iteration, it provides for a more thorough exploration of the data set, resulting in a solution of increased accuracy.



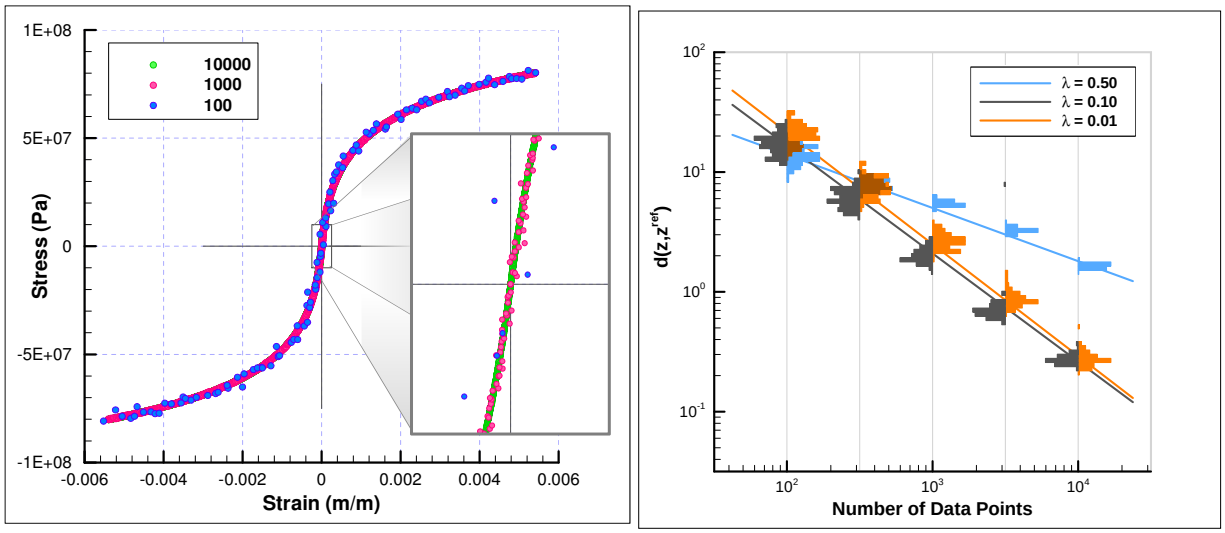
(a) (b)

Figure 3.5: Truss test case. a) Error in the data driven solution relative to the reference solution as a function of λ and data set size. b) Convergence to the reference solution with increasing data set size.

Further evidence of this annealing speed *vs.* accuracy trade-off is collected in Fig. 3.5. Thus, Fig. 3.5a shows the error in the data driven solution relative to the reference solution as a function of λ and data set size. As may be seen from the figure, the data driven solution is relatively insensitive to λ for small, or coarse, data sets. This lack of sensitivity owes to the fact that, by virtue of the coarseness of the data set, the simulated-annealing iteration leads to identical, or nearly identical, local data cluster regardless of the value of λ. By contrast, the range of possible limiting local data clusters increases with the size of the data set. Under these conditions, a conservative annealing schedule is more effective at identifying an optimal, or nearly-optimal, local data set cluster, at an attendant improvement in the accuracy of the solution. Fig. 3.5a also illustrates the beneficial effect of performing a distance-minimizing iteration after quenching (grey symbols) *vs.* stopping the iteration upon quenching (orange symbols). Fig. 3.5b shows the rates of convergence achieved as a function of λ and the size of the data set. The theorem supporting Propostion 1 in Chapter 2 shows that, for the data sets under consideration, the rate of convergence of distance-minimizing Data Driven solutions with respect to data set size is linear. Fig. 3.5b suggests that the same rate of convergence is achieved asymptotically by the max-ent Data Driven solutions for sufficiently small λ.

3.5.2 Uniform convergence of a noisy data set towards a classical material model

Next we consider data sets that, while uniformly convergent to a material curve in phase space, include noise in inverse proportion to the square root of the data set size. To construct a data set consistent with this aim, points are first generated directly from the material curve so that the metric distance between the points is constant. This first sample then has noise added independently pointwise according to a capped normal distribution in both the strain and stress axes with zero mean and standard deviation in inverse proportion to the square root of the data set size. The resulting data sets converge uniformly to the limiting material curve with increasing number of data points. Fig. 3.6a illustrates the data sets thus generated when the limiting model is as shown in Fig. 3.2b.



(a) (b)

Figure 3.6: Truss test case. a) Random data sets generated according to capped normal distribution centered on the material curve of Fig. 3.2b with standard deviation in inverse proportion to the square root of the data set size. b) Convergence with respect to data set size of error histograms generated from 100 material set samples.

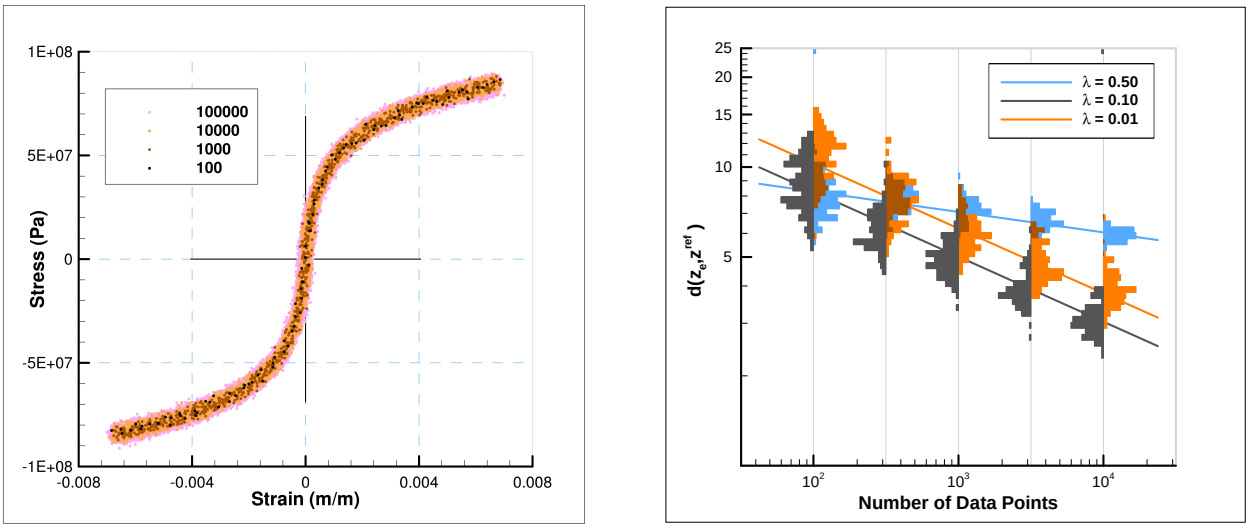
A convergence plot of error *vs.* data set size in shown in Fig. 3.6b, with error defined as the distance between the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 100 randomly generated data set samples. We again recall from Chapter 2, given the capped structure of the data sets under consideration, distance-minimizing Data Driven solutions converge to the limiting classical solution as $N^{-1/2}$, with

N the size of the data set. Surprisingly, an analysis of Fig. 3.6b suggests that, for sufficiently small λ, the max-ent Data Driven solutions converge as N^{-1} instead, i. e., they exhibit a linear rate of convergence with the data set size.

The random sampling of the data sets also raises questions of convergence in probability. It is interesting to note from Fig. 3.6b that both the mean error and the standard deviation of the error distribution converge to zero with increasing data set size. As already noted, the mean error exhibits a linear rate of convergence. The roughly constant width of the error histograms in log-log coordinates, suggests that the standard deviation of the error also converges to zero linearly with increasing data set size. These two observations together suggest that the error distribution obtained from a capped normal sampling of a material reference curve converges with sample size to the Dirac distribution centered at zero in both mean and in mean square, hence in probability [18].

3.5.3 Random data sets with fixed distribution about a classical material model

A different convergence scenario arises in connection with random material behavior described by a *fixed* probability measure μ in phase space. Specifically, given a set E in phase space, $\mu(E)$ is the probability that a fair test return a state $z \in E$. By virtue of the randomness of the material behavior, the solution becomes itself a random variable. We recall that the constraint set C is the set of states z in phase space that are compatible and in equilibrium. When the material behavior is random and is characterized by a probability measure μ in phase space, the solution must be understood in probabilistic terms and may be identified with the conditional probability $\mu \llcorner C$ of μ conditioned to C. The corresponding question of convergence then concerns whether the distribution of Data Driven solutions obtained by sampling μ by means of data sets of increasing size converges in probability to $\mu \llcorner C$.



(a) (b)

Figure 3.7: Truss test case. a) Random data sets generated according to normal distribution centered on the material curve of Fig. 3.2b with constant standard deviation independent of the data set size. b) Convergence with respect to data set size of error histograms generated from 100 material set samples.

While a rigorous treatment of convergence in probability is beyond the scope of this thesis, we may nevertheless derive useful insights from numerical tests. We specifically assume that μ is the cartesian product of member-wise measures μ_e characterizing the material behavior of each bar e. Specifically, given a set E_e in the phase space of member e, $\mu_e(E_e)$ is the probability that a fair test of member e return a state $z_e \in E_e$. In accordance with this representation, in calculations we generate data sets member-wise from a zero-mean normal distribution that is no longer capped and whose standard deviation is held constant. Fig. 3.7a illustrates the data sets thus generated when the base model is as shown in Fig. 3.2b.

Since the probability measure μ_e is generated by *adding* zero-mean normal random displacements to the base model in phase space, and since the constraint set C is *linear*, the conditional probability $\mu \llcorner C$ is itself centered on the base model. Hence, its mean value $\bar{z}$ necessarily coincides with the classical solution. This property is illustrated in Fig. 3.7b, which shows a convergence plot of error *vs.* data set size, with error defined as the distance between the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 100 randomly generated data set samples. As may be seen from the figure, the mean value of the histograms converges to zero with data set size, which is indicative of convergence in the

mean of the sampled max-ent Data Driven solutions. The rate of convergence of the mean error is computed to be of the order of 0.22. Interestingly, this rate of convergence is considerably smaller than the linear convergence rate achieved for the capped normal noise distributions considered in the preceding section. The slower rate of convergence may be attributable to the wider spread of the data about its mean, though the precise trade-off between convergence and uncertainty remains to be elucidated rigorously. Finally, we note from Fig. 3.7b that, as in the case of capped normal noise, an overly fast annealing schedule results in a degradation of the convergence rate.

3.6 Summary and discussion

We have formulated a Data Driven Computing paradigm, which we have termed max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing of the type developed in Chapter 2 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 problem consists of the minimization of a suitably-defined free energy over phase space subject to compatibility and equilibrium constraints. The problem is non-standard in the sense that the relevant Data Driven free energy is defined jointly over driving forces and fluxes. The distance-minimizing Data Driven schemes of Chapter 2 are recovered in the limit of zero temperature. We have also developed a simulated annealing solver that delivers the solution through a suitably-defined quenching schedule. Finally, we have presented selected numerical tests that establish the good convergence properties of the max-ent Data Driven solutions and solvers.

We conclude by framing Data Driven Computing within the context of past and present efforts to automate the connection between data and material models and expounding on how Data Driven Computing differs from said efforts. We also point out a number of possible enhancements of the approach that define worthwhile directions of further research.

3.6.1 Irreducibility to classical material laws

As already noted, the Data Driven free energy (3.7) and the associated problem (3.8) are non-standard. Thus, if $z = (\epsilon, \sigma)$, with ϵ the collection of local states of the system and σ the corresponding fluxes, the classical free energy is a function of state of the form $A(\epsilon, \beta)$, i. e., it is a

function of the driving forces and temperature, and the fluxes follow as $\sigma = \partial_\epsilon A(\epsilon, \beta)$. The corresponding classical problem consists of minimizing $A(\epsilon, \beta)$ with respect to the driving forces ϵ subject to compatibility constraints. Correspondingly, the classical Gibbs energy is a function of state of the form $G(\sigma, \beta)$, i. e., a function of the fluxes and temperature, and the driving forces follow as $\epsilon = \partial_\sigma G(\sigma, \beta)$. The corresponding classical problem consists of minimizing $G(\sigma, \beta)$ with respect to the fluxes σ subject to equilibrium constraints. By contrast, here the relevant free energy (3.7) is a function $F(z, \beta)$ defined over the entire phase space, i. e., is a joint function of the driving forces and fluxes. The corresponding Data Driven problem consists of minimizing $F(z, \beta)$ with respect to z subject to compatibility and equilibrium constraints simultaneously. Of course, it is possible to define an effective free energy through a partial minimization of $F(z, \beta)$ with respect to the fluxes, i. e.,

$$A(\epsilon, \beta) = \min \{F\big((\epsilon, \sigma), \beta\big), (\epsilon, \sigma) \in C\}, \tag{3.62}$$

with $A = +\infty$ is no minimizer exists. The corresponding effective classical problem then consists of minimizing $A(\epsilon, \beta)$ with respect to the driving forces ϵ. Likewise, it is possible to define an effective Gibbs energy through a partial minimization of $F(z, \beta)$ with respect to the driving forces, i. e.,

$$G(\sigma, \beta) = \min \{F\big((\epsilon, \sigma), \beta\big), (\epsilon, \sigma) \in C\}, \tag{3.63}$$

with $G = +\infty$ is no minimizer exists. The corresponding effective classical problem then consists of minimizing $G(\sigma, \beta)$ with respect to the fluxes σ. However, we note that these effective energies are *global* and do not correspond to a classical *local* material law in general. For instance, consider a finite-dimensional problem, such as the truss example developed in the foregoing, with compatibility and equilibrium constraints that can be satisfied identically through the representations

$$\epsilon = Bu, \tag{3.64a}$$

$$\sigma = A\varphi, \tag{3.64b}$$

where u is a displacement vector, B is a discrete strain operator, φ is an Airy potential vector and A is a discrete Airy operator, with the properties

$$B^T A = 0, \tag{3.65a}$$

$$A^T B = 0, \tag{3.65b}$$

which identify B^T and A^T as the discrete equilibrium and compatibility operators, respectively. In this representation, the Data Driven problem becomes

$$F(Bu, A\varphi, \beta) \to \min! \tag{3.66}$$

The corresponding Euler-Lagrange equations are

$$A^T \frac{\partial F}{\partial \sigma}(Bu, A\varphi, \beta) = 0, \tag{3.67a}$$

$$B^T \frac{\partial F}{\partial \epsilon}(Bu, A\varphi, \beta) = 0, \tag{3.67b}$$

which represent the discrete compatibility and equilibrium equations, respectively. Evidently, it is now possible to eliminate the Airy potential vector φ using the compatibility equations (3.67a) to define a reduced equilibrium problem in the displacement vector u, or, alternatively, eliminate the displacement vector u using the equilibrium equations (3.67b) to define a reduced compatibility problem in the Airy potential vector φ. However, these reduced problems remain non-classical in that they are *non-local*, i. e., they do not correspond to any local member-wise material law in general.

3.6.2 Material Informatics

There has been extensive previous work focusing on the application of Data Science and Analytics to material data sets. The field of Material Informatics (cf., e. g., [85, 81, 23, 82, 80, 84, 83, 22, 47, 49]) uses data searching and sorting techniques to survey large material data sets. It also uses machine-learning regression [19, 93] and other techniques to identify patterns and correlations in the data for purposes of combinatorial materials design and selection. These approaches represent an application of standard sorting and statistical methods to material data sets. While efficient at looking up and

sifting through large data sets, it is questionable that any real epistemic knowledge is generated by these methods. What is missing in Material Informatics is an explicit acknowledgement of the field equations of physics and their role in constraining and shaping material behavior. By way of contrast, such field equations play a prominent role in the Data Driven Computing paradigm developed in the present work.

3.6.3 Material identification

There has also been extensive previous work concerned with the use of empirical data for parameter identification in prespecified material models, or for automating the calibration of the models. For instance, the Error-in-Constitutive-Equations (ECE) method is an inverse method for the identification of material parameters such as the Young's modulus of an elastic material [41, 16, 35, 17, 72, 96, 10, 77, 68, 75]. While such approaches are efficient and reliable for their intended application, namely, the identification of material parameters, they differ from Data Driven Computing in that, while material identification schemes aim to determine the parameters of a prespecified material law from experimental data, Data Driven Computing dispenses with material models altogether and uses fundamental material data directly in the formulation of initial-boundary-value problems and attendant calculations thereof.

3.6.4 Data repositories

A number of repositories are presently in existence aimed at data-basing and disseminating material property data, e. g., [4, 2, 1, 3]. However, it is important to note that the existing material data repositories archive parametric data that are specific to prespecified material models. For instance, a number of repositories rely on parameterizations of standard interatomic potentials, such as the embedded-atom method (EAM), and archive data for a wide range of materials systems. Evidently, such data are strongly biased by—and specific to—the assumption of a specific form of the interatomic potential. By way of contrast, Data Driven Computing is based on fundamental, or *model-free*, material data only. Thus, suppose that the problem of interest is linear elasticity. In this case, the field equations are the strain-displacement and the equilibrium relations, and the local states are described by a strain tensor and a corresponding stress tensor. It thus follows that, in this case, model-free fundamental data consists of points in strain-stress space, or *phase space*.

By relying solely on fundamental data, Data Driven Computing requires no *a priori* assumptions regarding particular forms, and parameterizations thereof, of material models.

3.6.5 Implementation improvements

This chapter has focused on a particular definition of the annealing schedule as a means to implementing the new class of max-ent Data Driven solvers. It remains easily within the bounds of expectation that improvements in the schedule definition could lead to reductions in the number of iterations and improvements in annealing convergence rates. A number of other implementation improvements are equally worthy of examination. At present, sums over entire data sets for each material point were calculated without simplification or truncation. However, early stages of the annealing schedule could easily be performed on subsampled or summarized data sets due to the nonlocal nature of the calculations. Late stages of the annealing schedule could easily truncate sums over the data set through the use of cutoff radii pegged to $\beta^{-1/2}$. These and other considerations are likely to play an important role in the progression towards efficient and scalable implementations of the method.

3.6.6 Data coverage, sampling quality, adaptivity

Data Driven solvers provide, as a by-product, useful information regarding data coverage and sampling quality. Specifically, suppose that z is a Data Driven solution and z_e is the corresponding local state at material point e. Then, the distance $d_e(z_e, E_e)$ supplies a measure of how well the local state z_e is represented within the local material data set E_e. For any given material data set, a certain spread in the values of $d_e(z_e, E_e)$ may be expected, indicating that certain local states in a solution are better sampled than others. Specifically, local states with no nearby data points result in high values of $d_e(z_e, E_e)$, indicative of poor coverage by the material data set. Thus, the analysis of the local values $d_e(z_e, E_e)$ of the distance function provides a means of improving material data sets adaptively for particular applications. Evidently, the optimal strategy is to target for further sampling the regions of phase space corresponding to the local states with highest values of $d_e(z_e, E_e)$. In particular, local states lying far from the material data set, set targets for further testing. In this manner, the material data set may be adaptively expanded so as to provide the best possible coverage for a particular application.

3.6.7 Data quality, error bounds and confidence

Not all data are created equal, some data are of higher quality than others. In general, it is important to keep careful record of the pedigree, or ancestry, of each data point and to devise metrics for quantifying the level of confidence that can be placed on the data [74]. The confidence level in a material data point z_i can be quantified by means of a confidence factor $c_i \in [0, 1]$, with $c_i = 0$ denoting no confidence and $c_i = 1$ denoting full confidence. The weighting of the data points can then be modified to

$$p_i(z, \beta) = \frac{c_i}{Z(z, \beta)} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}, \tag{3.68a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} c_i \mathrm{e}^{-\beta d^2(z, z_i)}, \tag{3.68b}$$

which effectively factors the confidence factors into the calculations. In addition, material data obtained through experimental measurements often comes with error bounds attached. The standard error of a measurement of mean z_i is normally identified with its standard deviation s_i. In such cases, assuming the distribution of measurements to be Gaussian we obtain the distribution of weights

$$p_i(z, \beta) = \frac{1}{Z(z, \beta)} \mathrm{e}^{-1/2(s_i^2 + 1/2\beta)^{-1} d^2(z, z_i)}, \tag{3.69a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} \mathrm{e}^{-1/2(s_i^2 + 1/2\beta)^{-1} d^2(z, z_i)}. \tag{3.69b}$$

Again, this simple device effectively factors the experimental error bounds into the calculations.

Chapter 4

Dynamics Constraints as Applied to Different Schemes

This work is available as a preprint article [56] and has since been submitted for publication [59]. It is presented here with only small modifications.

4.1 Introduction

As we transition into an era of data generation and collection, constitutive relations will increasingly come to be characterized by information sets that are *data rich* throughout the regimes of interest. In this new environment, where inference is no longer required, empirical summaries will be necessarily less rich than the data upon which they were based. In these circumstances, modeling finds itself unable to take full advantage of the increasingly large data sets. Ultimately, the assumptions of a model become a restriction on the ability of a calculation to reproduce observed behavior. This lack of predictiveness then leads to unresolvable modeling errors that detract from the quality of the solution. The question then becomes how to move scientific computing beyond the modeling paradigm and have it operate directly on the supplied data sets. This question is strongly reminiscent of the new field of *Data Science*. In its most general form, Data Science is the extraction of *knowledge* from large volumes of unstructured data [8, 9, 7, 14]. It uses analytics, data management, statistics and machine learning to derive mathematical models for subsequent use in decision making. Data Science already provides classification methods capable of processing source data directly into query answers in non-STEM problems. In a similar vein, there has been extensive previous work focusing on the application of Data Science and Analytics to material data

sets. The field of Material Informatics (cf., e. g., [85, 81, 23, 82, 80, 84, 83, 22, 47, 49]) uses data searching and sorting techniques to survey large material data sets. It also uses machine-learning regression [19, 93] and other techniques to identify patterns and correlations in the data for purposes of combinatorial materials design and selection. However, what is missing in Material Informatics is an explicit acknowledgement of the field equations of physics and their role in constraining and shaping material behavior. At its best, Material Informatics represents an application of standard sorting and statistical methods to material data sets.

While efficient at looking up and sifting through large data sets, it is questionable that any real epistemic knowledge is generated by these methods.

There has also been extensive previous work concerned with the use of empirical data for parameter identification in prespecified material models, or for automating the calibration of the models. For instance, the Error-in-Constitutive-Equations (ECE) method is an inverse method for the identification of material parameters such as the Youngs modulus of an elastic material [41, 16, 35, 17, 72, 96, 10, 77, 68, 75]. While such approaches are efficient and reliable for their intended application, namely, the identification of material parameters, they differ from Data Driven Computing, as understood here, in that, while material identification schemes aim to determine the parameters of a prespecified material law from experimental data, Data Driven Computing dispenses with material models altogether and uses fundamental material data directly in the formulation of initial-boundary-value problems and attendant calculations thereof.

It was shown in previous chapters that it is indeed possible to reformulate the classical boundary-value problems of mechanics directly in terms of material data alone, without recourse to material modeling, pre-analysis or pre-processing of the material data. In this Data Driven Computing paradigm, the compatibility and conservation laws are recognized as material-independent differential constraints. The Data Driven solution is then defined as the point of the constraint manifold that is closest to the material data set. In this manner, the solution is determined directly by the material data, without recourse to any modeling of the data. Chapter 2 demonstrated the distance-based paradigm to be well-posed with respect to uniformly convergence of the material-data set. The effect of outliers in the material data set can be further mitigated by means of maximum-entropy estimation and information theory in Chapter 3.

The present chapter is concerned with the extension of Data Driven computing to dynamics.

Distance-minimizing methods described in Chapter 2 are encompassed as a special case of the applied annealing schedule. Time is discretized using a variational time stepping scheme that is used to generalize the static equilibrium constraints used in previous work. Selected numerical tests are used to demonstrate the convergence properties of both distance minimizing and entropy maximizing data solvers.

The chapter is organized as follows. In Section 4.2, we review max-ent Data Driven solvers and the associated simulated annealing schedules needed for their implementation. In Section 4.3 we extend previous Data Driven solvers, concerned with quasistatic problems, to dynamics. In Section 4.4, we present numerical tests that assess the convergence properties of max-ent and distance minimizing Data Driven solutions with respect to uniform convergence of the material data set. We also demonstrate the performance of max-ent based Data Driven Computing when the material behavior itself is random, i. e., defined by a probability density over the phase space. This is followed by a qualitative discussion of performance for these methods, beyond the specifics of convergence. Finally, concluding remarks and opportunities for further development of the Data Driven paradigm are presented in Section 4.5.

4.2 Review of Data Driven schemes

A main task of scientific calculations is to resolve coupled field responses to boundary conditions. Constitutive relationships then define the nature of coupling between the related fields. The language here restricts itself to mechanics, but mechanics is itself a special case of potential field theory through which electrostatics, diffusion and others present a similar need for constitutive definitions. Continuing within mechanics, the relations of interest are the extensive kinematic and kinetic work conjugate fields, e.g. ε and σ. Individually these fields must satisfy *material independent* properties with strong constraints. Kinematic fields must satisfy compatibility, while kinetic fields conserve momentum to be consistent with known physical laws. The certainty with which such field constraints can be asserted stands in stark contrast to the *material dependent* constitutive model which typically relates the two fields. Such models must be informed by supplied data, whose summarization into a model is typically performed using ad-hoc empirical fits. These fits, while providing speed and the opportunity for the introduction of intuition and inference, simultaneously introduce a modeling error that influences computational conclusions in ways which are hard to characterize.

To move beyond modeling, we now focus on the material data sets upon which such a models are based. This data E exists as a finite point set in phase space Z, where an example from small deformation mechanics would express the set as $E = ((\varepsilon_i, \sigma_i), i = 1, \cdots, N)$. The discrete nature of the set would naturally confound constitutive strategies which rely upon making use of a characterized function form. If compatibility, equilibrium, and boundary conditions are represented by the constraint set C, a problem arises in the likely case where the combined constraints cannot be satisfied by couplings defined by the discrete data set, thus $E \cap C$ returns an empty set. What is sought then is a relaxation which continues to satisfy all the members of C while minimizing deviations from E through direct data references.

Initially in Chapter 2, work on Data Driven computing focused primarily on establishing and demonstrating of a new class of Data Driven solvers through the use of a distance minimizing argument. However, such solvers exhibit data-convergence for noisy sets only if the sequence of data sets converges to a graph in the phase space. The need to accommodate a finite band of data obviates the need for a *probabilistic* solution strategy which arbitrates on the relevance and importance of different data points based on proximity. *Cluster analysis* provides a means of incorporating the influence of data neighborhoods to allow data-convergence in the presence of deeper samplings of fixed distributions.

4.2.1 Data clustering

Data Driven solvers for noisy data were developed in Chapter 3, employing cluster analysis to make a new kind of data driven solver, robust to outliers and well suited to data sources with finite data bands. The foundations of cluster analysis have their roots in concepts provided by Information Theory, such as *maximum-entropy* estimation [52]. Specifically, we wish to quantify how well a point z in phase space is represented by a point z_i in a material data set $E = (z_1, \cdots, z_n)$. Equivalently, we wish to quantify the *relevance* of a point z_i in the material data set to a given point z in phase space. We measure the relevance of points z_i in the material data set by means of *weights* $p_i \in [0, 1]$ with the property

$$\sum_{i=1}^{n} p_i = 1. \tag{4.1}$$

We wish the ranking by relevance of the material data points to be *unbiased*. It is known from Information Theory that the most unbiased distribution of weights is that which maximizes *Shannons*

information entropy [89, 90, 91]. In addition, we wish to accord points distant from z less weight than nearby points. These competing objectives can be combined by introducing a Pareto weight $\beta \geq 0$. The optimal and least-biased distribution is given by the Bolzmann distribution[89, 13]:

$$p_i(z, \beta) = \frac{1}{Z(z, \beta)} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}, \tag{4.2a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}. \tag{4.2b}$$

The corresponding max-ent Data Driven solver now consists of minimizing the free energy

$$F(z, \beta) = -\frac{1}{\beta} \log Z(z, \beta), \tag{4.3}$$

over the constraint set C, i. e.,

$$z \in \operatorname{argmin}\{F(z', \beta),\ z' \in C\}. \tag{4.4}$$

For finite β, all points in the material data set influence the solution, but their corresponding weights diminish with distance to the solution. In particular, the addition of an outlier that is marginally closer to the constraint set C than a large cluster of material data points does not significantly alter the solution.

4.2.2 Fixed point iteration

Having defined the max-ent Data Driven problem of interest to be the minimization of the free energy $F(z)$ (4.3) over the constraint set C. The corresponding optimality condition is

$$\frac{\partial F}{\partial z}(z, \beta) \perp C, \tag{4.5}$$

where $\perp$ denotes orthogonality. Assuming

$$d(z, z') = |z - z'|, \tag{4.6}$$

with $|\cdot|$ the standard norm in $\mathbb{R}^n$, we compute

$$\frac{\partial F}{\partial z}(z,\beta) = \sum_{i=1}^{n} p_i(z,\beta)(z - z_i) = z - \sum_{i=1}^{n} p_i(z,\beta) z_i. \tag{4.7}$$

Inserting this identity into (4.5), we obtain

$$z - \sum_{i=1}^{n} p_i(z,\beta) z_i \perp C, \tag{4.8}$$

which holds if and only if

$$z = P_C \left(\sum_{i=1}^{n} p_i(z,\beta) z_i \right), \tag{4.9}$$

where P_C is the closest-point projection to C. For instance, if $C = \{f(z) = 0\}$ for some constraint function $f(z)$, (4.5) may be expressed as

$$\frac{\partial F}{\partial z}(z,\beta) = \eta \frac{\partial f}{\partial z}(z), \tag{4.10a}$$

$$f(z) = 0, \tag{4.10b}$$

where η is a Lagrange multiplier. We note that eq. (4.9) conveniently defines the following fixed-point iteration,

$$z^{(k+1)} = P_C \left(\sum_{i=1}^{n} p_i(z^{(k)},\beta) z_i \right). \tag{4.11}$$

The essential difficulty inherent to problem (4.5), or (4.10), is that, in general, the free energy function $F(\cdot,\beta)$ is strongly non-convex, possessing multiple wells centered at the data points in the material data set. Under these conditions, iterative solvers may fail to converge or may return a local minimizer, instead of the global minimizer of interest. We overcome these difficulties by recourse to *simulated annealing* [61].

4.2.3 Simulated annealing

The general idea of simulated annealing is to evolve the reciprocal temperature jointly with the fixed point iteration according to an appropriate annealing schedule, i. e., we modify (4.11) to

$$z^{(k+1)} = P_C \left(\sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)}) z_i \right). \tag{4.12}$$

An effective annealing schedule is obtained by selecting $\beta^{(k+1)}$ so as to ensure local contractivity of the fixed-point mapping. An appeal to contractivity in Chapter 3 suggests the schedule

$$\frac{1}{\beta^{(k+1)}} = \sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)}) |z_i - \bar{z}^{(k)}|^2, \tag{4.13}$$

with the initial reciprocal temperature $\beta^{(0)}$ chosen small enough that the mapping $g(\cdot, \beta^{(0)})$ is contractive everywhere. This further leads to an estimate for a convexifying β_0 with which to initialize the iteration,

$$\frac{1}{\beta^{(0)}} = \frac{1}{n} \sum_{i=1}^{n} |z_i|^2. \tag{4.14}$$

4.3 Application to dynamics

We illustrate the extension of max-ent Data Driven Computing to dynamical problems by means of the simple example of truss structures. Trusses are assemblies of articulated bars that deform in uniaxial tension or compression. Thus, conveniently, in a truss the material behavior of a bar e is characterized by a simple relation between the uniaxial strain ε_e and uniaxial stress σ_e in the bar. We refer to the space of pairs $z_e = (\varepsilon_e, \sigma_e)$ as the *phase space* of bar e. We assume that the behavior of the material of each bar $e = 1, \ldots, m$, where m is the number of bars in the truss, is characterized by—possibly different—local data sets E_e of pairs z_e, or *local states*. For instance, each point in the data set may correspond, e. g., to an experimental measurement. The global data set is then the cartesian product $E = \prod_{e=1}^{m} E_e$ of all local data sets.

The state $z_k = (z_e)_{e=1}^{m}$ of the truss at some time t_k is subject to the compatibility and equilibrium

constraints

$$\varepsilon_e = B_e u_k, \tag{4.15a}$$

$$\sum_{e=1}^{m} B_e^T w_e \sigma_{e,k} = f_k - M a_k, \tag{4.15b}$$

where u and a are the array of nodal displacements and accelerations, f is the array of applied nodal forces, the matrices $(B_e)_{e=1}^m$ encode the geometry and connectivity of the truss members, w_e is the volume of member e, and M is the mass matrix. In order to integrate the equations in time we proceed to discretize displacement u and its derivatives v and a in time using the Newmark algorithm

$$u_{a,k} = u_{a,k-1} + \Delta t v_{a,k-1} + \Delta t^2 \left(\left(\frac{1}{2} - \beta \right) a_{a,k-1} + \beta a_{a,k} \right), \tag{4.16a}$$

$$v_{a,k} = v_{a,k-1} + \Delta t \left(1 - \gamma \right) a_{a,k-1} + \gamma a_{a,k} \big), \tag{4.16b}$$

where β and γ are the Newmark parameters. In order to reduce these equations to an equivalent static problem, we introduce the Newmark predictors

$$u_{a,k}^{\text{pred}} = u_{a,k-1} + \Delta t v_{a,k-1} + \left(\frac{1}{2} - \beta \right) \Delta t^2 a_{a,k-1}, \tag{4.17a}$$

$$v_{a,k}^{\text{pred}} = v_{a,k-1} + (1 - \gamma) \Delta t a_{a,k-1}, \tag{4.17b}$$

whereupon the constraints can now be written

$$\varepsilon_e = B_e u_k, \tag{4.18a}$$

$$\sum_{e=1}^{m} B_e^T w_e \sigma_{e,k} = f_k - M \frac{u_{a,k} - u_{a,k}^{\text{pred}}}{\beta \Delta t^2}, \tag{4.18b}$$

with an associated update

$$a_{a,k} = \frac{u_{a,k} - u_{a,k}^{\text{pred}}}{\beta \Delta t^2}, \tag{4.19a}$$

$$v_{a,k} = v_{a,k}^{\text{pred}} + \gamma \Delta t a_{a,k}. \tag{4.19b}$$

We may metrize the local phase spaces of each member of the truss by means of Euclidean distances derived from the norms

$$|z_e|_e = \left(\mathbb{C}\varepsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\right)^{1/2}, \tag{4.20}$$

for some positive constant $\mathbb{C}$. We may then metrize the global state of the truss by means of the global norm

$$|z| = \Big(\sum_{e=1}^{m} w_e |z_e|_e^2\Big)^{1/2} = \left(\sum_{e=1}^{m} w_e\Big(\mathbb{C}\varepsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\Big)\right)^{1/2} \tag{4.21}$$

and the associated distance (4.6). For a truss structure, the point in C closest to a given point z^* in phase space follows from the stationarity condition

$$\delta\left\{\sum_{e=1}^{m} w_e\left(\frac{\mathbb{C}}{2}(B_e u_k - \varepsilon_e^*)^2 + \frac{\mathbb{C}^{-1}}{2}(\sigma_e - \sigma_e^*)^2\right) + \left(f - M\Big(\frac{u_k - u_k^{\text{pred}}}{\beta\Delta t^2}\Big) - \sum_{e=1}^{m} w_e B_e^T \sigma_e\right)^T \eta\right\} = 0, \tag{4.22}$$

where η is an array of Lagrange multiplier enforcing the equilibrium constraints. The corresponding Euler-Lagrange equations are

$$\sum_{e=1}^{m} w_e B_e^T \mathbb{C}\left(B_e u_k - \varepsilon_{e,k}^*\right) - M\frac{\eta}{\beta\Delta t^2} = 0, \tag{4.23a}$$

$$\mathbb{C}^{-1}(\sigma_{e,k} - \sigma_{e,k}^*) = B_e\eta, \tag{4.23b}$$

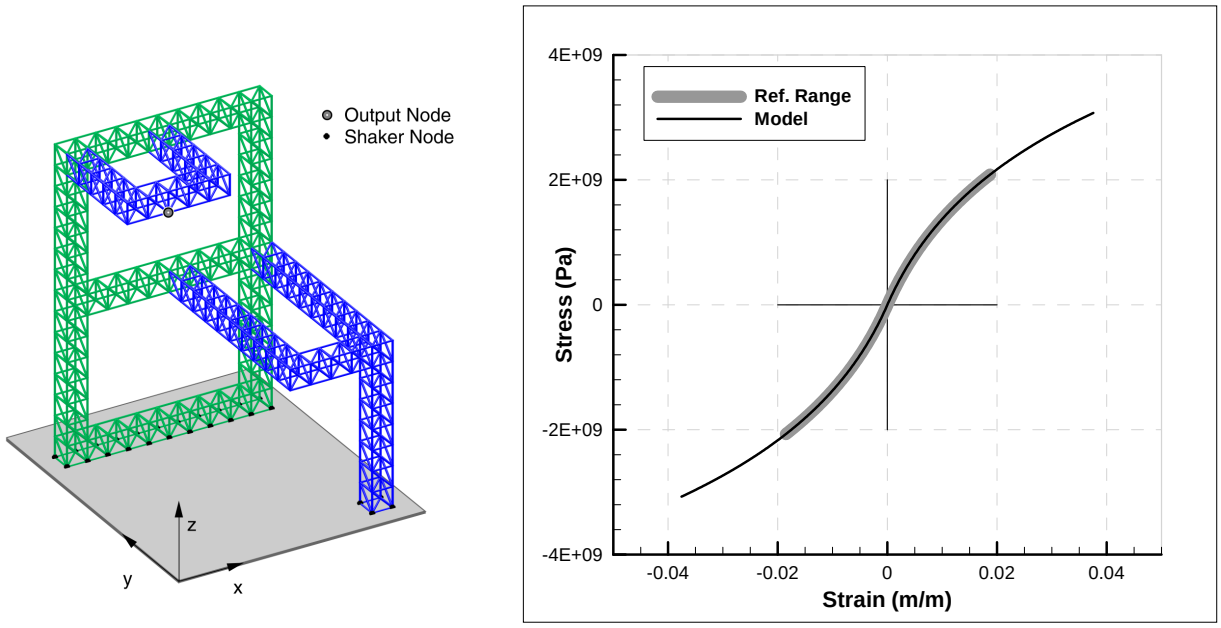
$$\sum_{e=1}^{m} w_e B_e^T \sigma_{e,k} = f_k - M\Big(\frac{u_k - u_k^{\text{pred}}}{\beta\Delta t^2}\Big), \tag{4.23c}$$

or

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) u_k = \sum_{e=1}^{m} w_e B_e^T \mathbb{C}\ \varepsilon_{e,k}^* + M\frac{\eta}{\beta\Delta t^2} \tag{4.24a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right)\eta = f_k - M\frac{u_k - u_k^{\text{pred}}}{\beta\Delta t^2} - \sum_{e=1}^{m} w_e B_e^T \sigma_e^*. \tag{4.24b}$$

which define two coupled truss equilibrium problems for the linear reference material of modulus $\mathbb{C}$.



(a) (b)

Figure 4.1: a) Geometry and boundary conditions of truss test case. b) Base material model with model sampling ranges superimposed.

4.4 Numerical tests

In calculations we consider the specific test case shown in Figure 4.1a. The truss contains 1,246 members and is supported as shown in the figure. An instantaneous sine excitation of 10 cycles was affected at the base-attached nodes over a time duration resolved with 300 steps. By way of reference, we consider the nonlinear stress-strain relation shown in 4.1b. A Newton-Raphson based solution with a consistently selected time integrator is readily obtained. The resulting range of states referenced over all the members of the truss in the course of the full time evolution are shown in Figure 4.1a superimposed on the stress-strain curve in order to visualize the coverage of phase space entailed by the reference solution.

To provide a measure of error that summarizes the performance of comparable time integration solutions requires a systematic way of comparing the solution across multiple timesteps without overweighting long-time phase error. Previous work of comparing transient finite element solutions

[76] is modified here to create such an analysis error metric,

$$\text{ERROR}^2 = \int_{t_1}^{t_f} \frac{1}{t^2} \sum_{e=1}^{m} w_e d^2(z, z^{\text{ref}}) dt, \tag{4.25}$$

where t_1 is the time after one step and t_f is the final time. The need for the lower limit of integration to start after the first time step arises from the singularity that would be created by a possible non-zero distance between reference and Data Driven under initial displacement conditions.

4.4.1 Annealing schedule

In this chapter we consider two independent annealing schedules to integrate time, both of which are described by algorithm 3. The first schedule sets $\beta_0 \to \infty$ and all subsequent steps continue to provide full weight to the nearest neighbor in the data set, thus making it consistent with a distance-minimizing scheme. This scheme was previously demonstrated for static mechanics problems in Chapter 2. The second schedule is based on maximum entropy. We specifically consider the case in which the behavior of each bar e is characterized by a local material data set $E_e = \{z_{i_e} = (\epsilon_{i_e}, \sigma_{i_e}) \in \mathbb{R}^2,\ i_e = 1, \ldots, n_e\}$, where n_e is the number of data points in E_e. The global data set is then the Cartesian product

$$E = E_1 \times \cdots \times E_m. \tag{4.26}$$

A typical point in such a data set is most convenient indexed as $z_{i_1 \ldots i_m}$, with $i_e = 1, \ldots, n_e$, $e = 1, \ldots, m$, instead of using a single index as in Section 4.2. The partition function (4.2b) then takes the form

$$Z(z, \beta) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \mathrm{e}^{-(\beta/2) \sum_{e=1}^{m} d^2(z_e, z_{i_e})}, \tag{4.27}$$

where the local distance is given by (4.20). Rearranging terms, (4.27) may be rewritten in the form

$$Z(z, \beta) = \prod_{e=1}^{m} \left(\sum_{i_e=1}^{n_e} \mathrm{e}^{-(\beta/2) d^2(z_e, z_{i_e})} \right) \equiv \prod_{e=1}^{m} Z_e(z_e, \beta), \tag{4.28}$$

and the total free energy evaluates to

$$F(z,\beta) = \sum_{e=1}^{m} \left(-\frac{1}{\beta} \log Z_e(z_e,\beta) \right) \equiv \sum_{e=1}^{m} F_e(z_e,\beta). \tag{4.29}$$

We note that the total free energy is additive with respect to the free energies $F_e(z_e,\beta)$ of the members. Finally, the Bolzmann distribution (4.2a) becomes

$$p_{i_1,\ldots,i_m}(z,\beta) = \prod_{e=1}^{m} \left(\frac{1}{Z_e(z_e,\beta)} \mathrm{e}^{-(\beta/2)d^2(z_e,z_{i_e})} \right) \equiv \prod_{e=1}^{m} p_{i_e}(z_e,\beta). \tag{4.30}$$

In the case of independent local material data sets, eq. (4.26), the bound (4.13), as seen in Chapter 3, specializes to

$$\frac{1}{\beta} < \sum_{e=1}^{m} \left(\sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta) d^2(\bar{z}_e, z_{i_e}) \right). \tag{4.31}$$

Continuing in the same vein as Chapter 3, we exploit this special structure and refine the bound by applying it at the local level, i. e., by requiring

$$\frac{1}{\beta_e} < \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta_e) d^2(\bar{z}_e, z_{i_e}), \tag{4.32}$$

$e = 1,\ldots,m$, where $1/\beta_e$ represent local temperatures. We can further define an annealing schedule by taking (4.32) as the basis for local temperature updates

$$\frac{1}{\beta_e^{(k+1)}} = \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta^{(k)}) d^2(\bar{z}_e^{(k)}, z_{i_e}), \tag{4.33}$$

with thermal equilibrium subsequently restored by setting the global temperature to

$$\frac{1}{\beta^{(k+1)}} = \sum_{e=1}^{m} \frac{w_e^{(k+1)}}{\beta_e^{(k+1)}}, \tag{4.34}$$

with appropriate weights $w_e^{(k+1)}$. In calculations, we specifically choose

$$w_e^{(k+1)} = \frac{\mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)},\beta_e^{(k)})}}{\sum_{e=1}^{m} \mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)},\beta_e^{(k)})}} = \frac{Z_e(\bar{z}_e^{(k)},\beta_e^{(k)})}{\sum_{e=1}^{m} Z_e(\bar{z}_e^{(k)},\beta_e^{(k)})}. \tag{4.35}$$

Finally, the initial estimate (4.14) corresponds to setting

$$p_{i_e}(z_e, \beta) = \frac{1}{n_e}, \tag{4.36}$$

whereupon (4.13) becomes

$$\frac{1}{\beta^{(0)}} = \sum_{e=1}^{m} \frac{1}{n_e} \left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e}) \right). \tag{4.37}$$

As a further control on the annealing rate we set

$$\beta^{(k+1)} = \lambda \tilde{\beta}^{(k+1)} + (1 - \lambda)\beta^{(k)}, \tag{4.38}$$

where $\tilde{\beta}^{(k+1)}$ is the result of applying the update (4.34) and λ is an adjustable factor.

Algorithm 3 Data Driven solver (one time step)

Require: Local data sets $E_e = \{z_{i_e},\, i_e = 1,\dots,n_e\}$, B-matrices $\{B_e,\, e = 1,\dots,m\}$, $k = 1$, initial displacements and velocities, force vector f, parameter λ.

1) Compute predictors u^{pred} and v^{pred} using eq. (4.17)

2) Initialize data iteration. Set $j = 0$, compute

$$\bar{z}_e^{(0)} = z_e^{(0)} = \frac{1}{n_e}\sum_{i_e=1}^{n_e} z_{i_e}, \quad \frac{1}{\beta^{(0)}} = \sum_{e=1}^{m}\frac{1}{n_e}\left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e})\right). \tag{4.39}$$

3) Calculate data associations and precalculate for convexity estimate:

for all $e = 1,\dots,m$ **do**

3.1) Set $c_{i_e}^{(j)} = \exp\big(-\beta^{(j)} d^2(z_e^{(j)}, z_{i_e})\big),\ i_e = 1,\dots,n_e$.

3.2) Set $Z_e^{(j)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(j)}$.

3.3) Set $p_{i_e}^{(j)} = c_{i_e}^{(j)}/Z_e^{(j)},\ i_e = 1,\dots,n_e$.

3.4) Set $\bar{z}_e^{(j)} = \sum_{i_e=1}^{n_e} p_{i_e}^{(j)} z_{i_e}$.

3.5) Set $D_e^{(j)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(j)} d^2(\bar{z}_e^{(j)}, z_{i_e})$

end for

4) Solve:

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) u = \sum_{e=1}^{m} w_e B_e^T \mathbb{C}\, \varepsilon_e^{(j)} + M\frac{\eta}{\beta\Delta t^2} \tag{4.40a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) \eta = f - M\frac{u - u^{\text{pred}}}{\beta\Delta t^2} - \sum_{e=1}^{m} w_e B_e^T \sigma_e^{(j)} \tag{4.40b}$$

for u and η.

5) Progress Schedule:

5.1) Set

$$\tilde{\beta}^{(j+1)} = \left(\frac{\sum_{e=1}^{m} D_e^{(j)}}{\sum_{e=1}^{m} Z_e^{(j)}}\right)^{-1}. \tag{4.41}$$

5.2) Set $\beta^{(j+1)} = (1-\lambda)\beta^{(j)} + \lambda\tilde{\beta}^{(j+1)}$.

6) Compute local states $z_{e,j}$:

for all $e = 1,\dots,m$ **do**

$$\varepsilon_e^{(j+1)} = B_e u^{(j+1)}, \qquad \sigma_e^{(j+1)} = \bar{\sigma}_e^{(j+1)} + \mathbb{C} B_e \eta^{(j+1)} \tag{4.42}$$

end for

7) Test for convergence and cycle the time or data iteration:

if $\{z_e^{(j+1)} = z_e^{(j)},\ e = 1,\dots,m\}$ **then**

exit

else

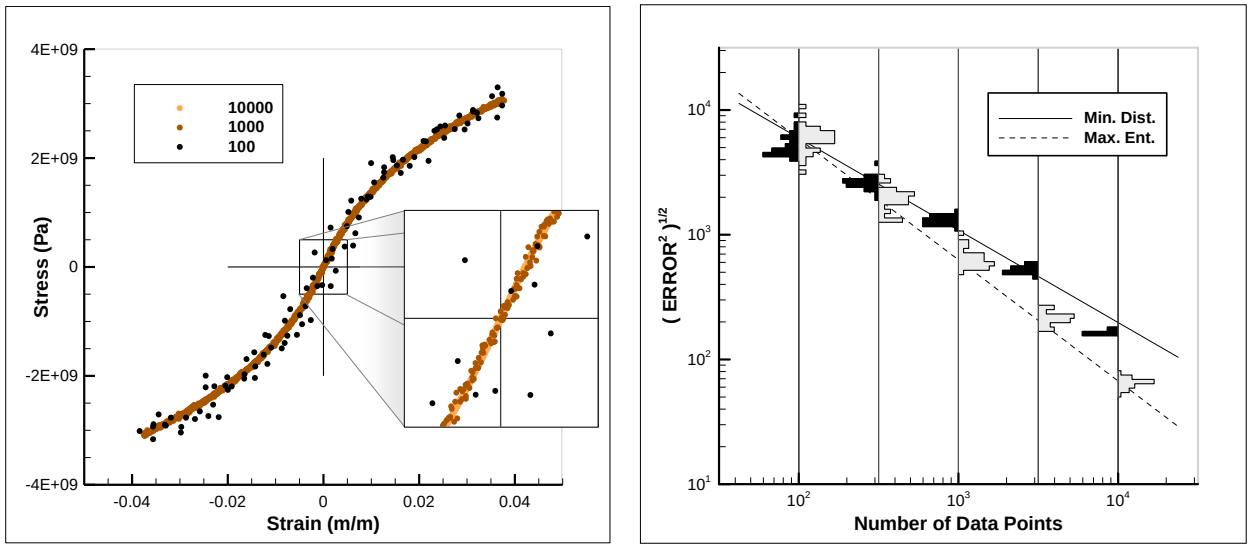
$j \leftarrow j + 1$,

goto (3).

end if

4.4.2 Uniform convergence of a noisy data set towards a classical material model

Next we consider data sets that, while uniformly convergent to a material curve in phase space, include noise in inverse proportion to the square root of the data set size. To construct a data set consistent with this aim, points are first generated directly from the material curve so that the metric distance between the points is constant. This first sample then has noise added independently pointwise according to a capped normal distribution in both the strain and stress axes with zero mean and standard deviation in inverse proportion to the square root of the data set size. The resulting data sets converge uniformly to the limiting material curve with increasing number of data points. Figure 4.2a illustrates the data sets thus generated when the limiting model is as shown in Figure 4.1b.



(a) (b)

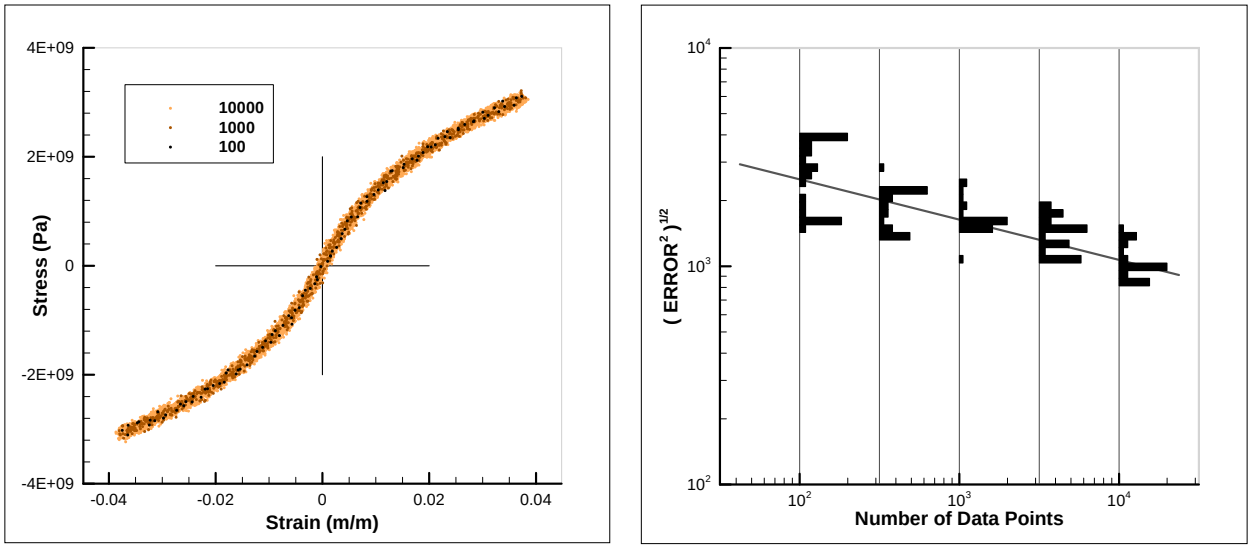
Figure 4.2: Truss test case. a) Random data sets generated according to capped normal distribution centered on the material curve of Figure 4.1b with standard deviation in inverse proportion to the square root of the data set size. b) Convergence with respect to data set size of error histograms generated from 30 material set samples.

A convergence plot of error *vs.* data set size in shown in Figure 4.2b, with error metric comparing the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 30 randomly generated data set samples. We see from Chapter 2 that, given the capped structure of the data sets under consideration, distance-minimizing Data Driven solutions converge to the limiting classical solution as $N^{-1/2}$, for the the data set size N.

An analysis of Figure 4.2b suggests that distance-minimizing schemes convergence rate of $N^{-1/2}$, which is consistent with methods employed for static analysis. Similarly, max-ent Data Driven solutions converge with a linear rate with respect to the data set size seen for static analysis in Chapter 3.

4.4.3 Random data sets with fixed distribution about a classical material model

A different convergence scenario arises in connection with random material behavior described by a *fixed* probability measure μ in phase space. Specifically, given a set E in phase space, $\mu(E)$ is the probability that a fair test return a state $z \in E$. By virtue of the randomness of the material behavior, the solution becomes itself a random variable. We recall that the constraint set C is the set of states z in phase space that are compatible and in equilibrium. When the material behavior is random and is characterized by a probability measure μ in phase space, the solution must be understood in probabilistic terms and may be identified with the conditional probability $\mu \llcorner C$ of μ conditioned to C. The corresponding question of convergence then concerns whether the distribution of Data Driven solutions obtained by sampling μ by means of data sets of increasing size converges in probability to $\mu \llcorner C$.



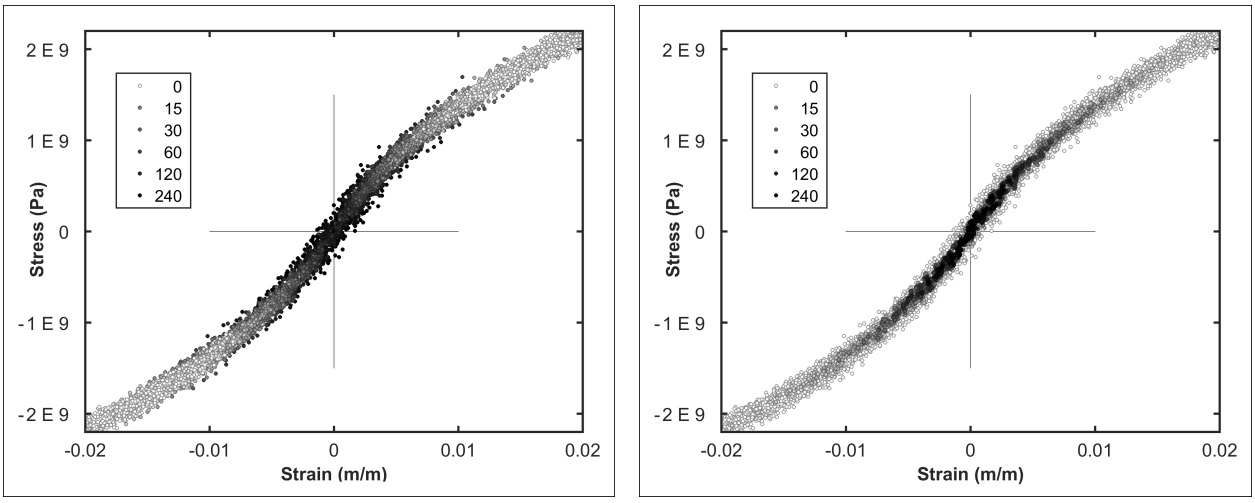
(a) (b)

Figure 4.3: Truss test case. a) Random data sets generated according to normal distribution centered on the material curve of Figure 4.1b with constant standard deviation independent of the data set size. b) Convergence with respect to data set size of error histograms generated from 30 material set samples.

While a rigorous treatment of convergence in probability is beyond the scope of this paper, we may nevertheless derive useful insights from numerical tests. We specifically assume that μ is the cartesian product of member-wise measures μ_e characterizing the material behavior of each bar e. Specifically, given a set E_e in the phase space of member e, $\mu_e(E_e)$ is the probability that a fair test of member e return a state $z_e \in E_e$. In accordance with this representation, in calculations we generate data sets member-wise from a zero-mean normal distribution that is no longer capped and whose standard deviation is held constant. Figure 4.3a illustrates the data sets thus generated when the base model is as shown in Figure 4.1b.

Since the probability measure μ_e is generated by *adding* zero-mean normal random displacements to the base model in phase space, and since the constraint set C is *linear*, the conditional probability $\mu \llcorner C$ is itself centered on the base model. Therefore, its mean value $\bar{z}$ necessarily coincides with the classical solution. This property is illustrated in Figure 4.3b, which shows a convergence plot of error *vs.* data set size, with error defined as the distance between the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 30 randomly generated data set samples. As may be seen from the figure, the mean value of the histograms converges to zero with data set size, which is indicative of convergence in mean of the sampled max-ent Data Driven solutions. The rate of convergence of the mean error is computed to be of the order of 0.19. Interestingly, this rate of convergence is considerably smaller than the linear convergence rate achieved for the capped normal noise distributions considered in the preceding section and also consistent with previous results of Chapter 3. The slower rate of convergence may be attributable to the wider spread of the data about its mean, though the precise trade-off between convergence and uncertainty remains to be elucidated rigorously.

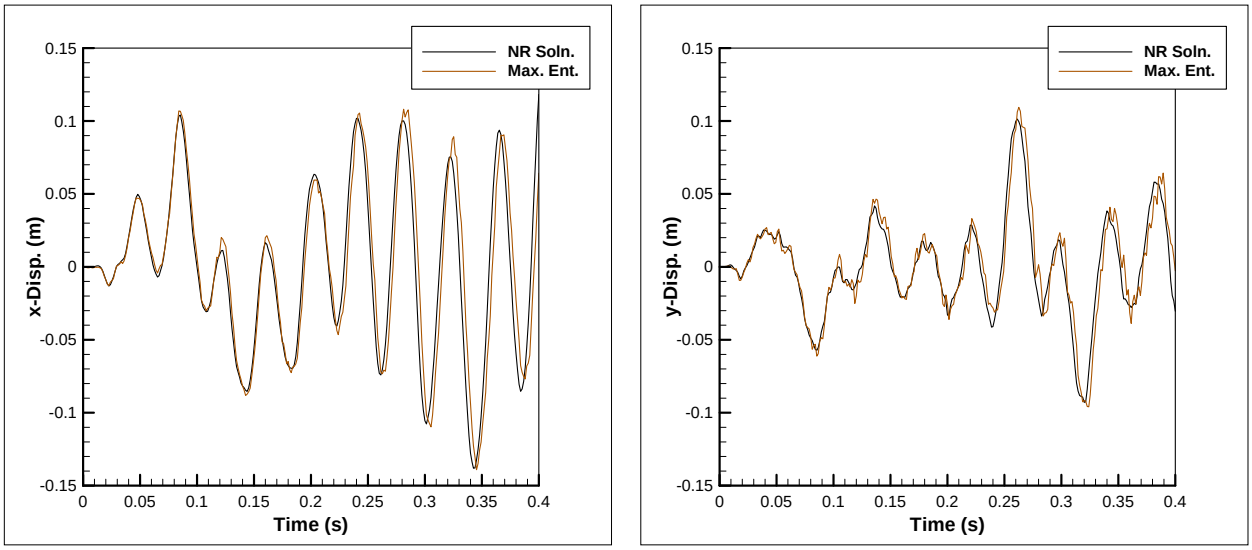
4.4.4 General performance characteristics



(a) (b)

Figure 4.4: Data set shaded by selection frequency for a) the distance minimizing and b) entropy maximizing selection schemes.

Even in the presence of noiseless data sets, the max-ent solutions are uniformly seen to be improvements on distance-minimizing solutions. These significant improvements arise out of the propensity for distance-minimizing schemes to become trapped in local minima semi-adjacent to true minimizers. In turn these local data selection error accumulate with successive time integration. The benefits of entropy maximizing solutions become especially apparent in Figure 4.4 where the 10,000 point data set is shaded based on the number of times the various data elements were referenced in the 300 step time solution. The distance-minimizing scheme shown in Figure 4.4a demonstrates how the algorithm not only allows for the selection of outliers, but how in some cases it *favors* the selection of outliers. The only region where outliers are not favored is near the point $z = (0, 0)$ where the elements are initialized. And while both methods were shown to converge, Figure 4.4 illustrates how the clustering argument made the max-ent solver robust to noisy data inputs, while the distance minimizing methods require the data converge to a graph in the phase space. Figure 4.5 shows how under the clustering argument the time history of displacement maintains a remarkable fidelity to the reference displacement history for the output node.



(a) (b)

Figure 4.5: Max-ent displacement solutions for geometry and boundary conditions seen in Figure 4.1a solved using the data set shown in Figure 4.4 for a) the x-direction displacement and b) the y-displacement.

4.5 Summary and discussion

We have applied methods of Data Driven Computing paradigm, including both distance-minimizing and max-ent schemes for Data Driven Computing, to a new set of time dependent problems. We then presented selected numerical tests that establish the good convergence properties for the implemented solvers. Both distance-minimizing and max-ent solutions were shown to converge as the sequence of sets converges to an underlying model. Max-ent solutions were additionally shown to be robust to outliers and converge as the sequence of data sets converged to a fixed distribution. Beyond the specific context of improved convergence rates and conditions, max-ent solvers were also shown to have much more efficacy for step driven transient solutions because of accumulated improvements of solutions over the time domain.

An essential aspect of the Data Driven paradigm is that the space of fundamental, or model-independent, data where material data take values is determined unambiguously by the compatibility and conservation laws. This reliance on fundamental, or model-independent, material data is an essential difference with the existing Data Repositories, e. g., [4, 2, 1, 3], which archive parametric

data that are specific to prespecified material models. Fundamental data is *fungible*, i. e., data that is raised for one purpose can be used for another. Fundamental data is also *blendable*, i. e., material data from different sources can be blended together into a single material data set. In particular, fundamental data repositories can be publicably editable, which opens up a new and potentially far-reaching way of pooling and distributing material data.

This paper has focused on re-implementing the annealing schedules explored in Chapters 2 and 3 as a means to demonstrating a new class of transient Data Driven solvers. As in previous implementations, we have made no effort to speed up the implementation of the respective schedules. Thus, there remain a number of improvements suggested in previous chapters, e. g., summarizing data sets, efficient range searches and radial cutoffs for summation. Time integration as affected through time stepping yields additional improvements, not implemented here, which would dramatically improve numerical performance. Specifically, previous time steps could be used to inform initialization values for an annealing process, analogous to similar strategies seen in non-linear time integration methodologies. Such initializations would allow schedules used for annealing to initiate from a $\beta^{(0)}$ larger than one which guarantees convexity over the whole data set.

Chapter 5

Conclusion

Significant text from this chapter is taken from [58].

5.1 Results summary

The work of this thesis cumulatively presents the opening stages of a new developement in the performance of scientific calculations. Data has long been used to indirecty inform numerical preditions through the introduction of constitutive models, these new methods create a new opportunity to directly associate constitutive data and the solutions they supply. The work of Chapter 2 provided a first description and example of what properties a numerical analysis could be described as being Data Driven Computing. The elegant distance-minimizing solvers then acted as an efficient vehicle for such discussions and were shown to converge for sequences of data sets which converge to a graph in the phase space. Chapter 3 then went on to develope max-ent based cluster argument to inform a method capable of converging in the presense of data sets which contain persistent outliers. These significant changes in solver properties allowed the solver to converge for sequences of data sets which converge to a data density distribution in the phase space. The static analyses used to demonstrate both methods were extended in Chapter 4 to include dynamics problems. The ability to modify the applied constraints to new classes of problems stands as a relevant example of the broad range of future applications to which Data Driven Computing could apply.

5.2 Method extensions to Data Driven Computing

In this section section we discuss the possible improvements to the new methods and applications explored in this thesis. Subsection 5.2.1 discusses the direct improvements on the methods thus far investigated so as to optimize the algorithms and annealing schedules for computational speed. These improvements can be largely framed as numerical accelereations or simplifications, max-ent solvers additionally have possible annealing schedule improvements. Moving beyond performance, Subsection 5.2.2 identifies Data Driven solver extensions which could incorporate data and data sources which contain descriptive confidence metrics. These improvements are a natural outgrowth of clustered annealing solvers. The remaining method improvement, discussed in Subsection 5.2.3, would be to incorporate intermediate Data Driven results to drive improved material response sampling.

5.2.1 Annealing schedule improvements

All methods thus far discussed exibit apperent opportunities for speed related improvements. Both distance-minimizing and max-ent based solvers must perform searches or calculations over the full data sets to progress towards selecting a final solution. Early stages of these evolution schedules could equivalently operate on sub-sampled or data-summary sets which would reduce the cost of search and calculation. Particularly in the case of max-ent solvers, this improvement would arise from the large convex neighborhood employed by the solvers in the beginning of the anneling schedule, which would depreciate the importance of any specific data member. Max-ent solvers could also benefit in the final stages of their schedule through the use of a cutoff radus to prevent unnecessary references to near-zero summation terms.

Beyond the apparent improvements to the numerical implementations already presented within this paper, there has also been no consideraton placed on optimizing the path of annealing schedules for max-ent based problems. These improvements could be framed has having an annealing scheudle where the λ defined in Chapter 3 and 4 is no longer constant, but instead dependent on step-to-step error or some aspect of how the data set is referenced. Finally the schedule used for dynamics in Chapter 4 is systematically conservative due to re-use of the schedule devised in Chapter 3. It would seem likely that schedules initialized from a previous time step could be initiated from a value for β_0 which is not convex over the entire data set.

5.2.2 Data quality, error bounds, confidence

In general, it is important to keep careful record of the pedigree, or ancestry, of each data point and to devise metrics for quantifying the level of confidence that can be placed on the data [74]. The confidence level in a material data point z_i can be quantified by means of a confidence factor $c_i \in [0, 1]$, with $c_i = 0$ denoting no confidence and $c_i = 1$ denoting full confidence. The weighting of the data points can then be modified to

$$p_i = \frac{c_i}{Z} \exp\left(-d^2(z, z_i)\right), \qquad Z = \sum_{i=1}^{n} c_i \exp\left(-d^2(z, z_i)\right),$$

which effectively factors the confidence factors into the calculations. In addition, material data obtained through experimental measurements often comes with error bounds attached. The standard error of a measurement of mean z_i is normally identified with its standard deviation s_i. In such cases, assuming the distribution of measurements to be Gaussian we obtain the distribution of weights:

$$p_i = \frac{1}{Z} \exp\left(\frac{-d^2(z, z_i)}{2s_i^2 + 1/\beta}\right), \qquad Z = \sum_{i=1}^{n} \exp\left(\frac{-d^2(z, z_i)}{2s_i^2 + 1/\beta}\right).$$

These simple analytical devices effectively factors the experimental error bounds and quality estimates directly into Data Driven solvers.

5.2.3 Data coverage, sampling quality, adaptivity

Data Driven solvers have the capacity, not just to make use of supplied data, but also suggest what additional data would be most useful in improving a characterized response. Recall that in general $z \in C$ represents a sampling of the Data Driven field value and $y \in E$ corresponds to data associations specific to the integration points. Solutions to Data Driven problems then naturally supply solutions z_e and y_e at each of the integration points e. The distance $d_e(z_e, E_e)$ then supplies a natural measure of how well the local state z_e is represented within the local material data set E_e. For any given material data set, a certain spread in the values of $d_e(z_e, E_e)$ may be expected, indicating that certain local states in a solution are better sampled than others. Specifically, local states with no nearby data points result in high values of $d_e(z_e, E_e)$, indicative of poor coverage by the material data set. Thus, the analysis of the local values $d_e(z_e, E_e)$ of the distance function provides a means of improving material data sets adaptively for particular applications. Evidently,

the optimal strategy is to target for further sampling the regions of phase space corresponding to the local states with highest values of $d_e(z_e, E_e)$. Local states lying far from the material data set become targets for further testing. In this manner, the material data set may be adaptively expanded to provide the best possible coverage for a particular application.

Some care in exercising the use of additional sampling in the presence of data clustering arguments to prevent a targeting bias. More specifically, it is expected that the introduction of additional data would suggest a strategy of "reheating" a solution and resuming an annealing process. Since clustering arguments cause regions of the phase space with denser data sampling to be favored, any additional material sampling then needs to populate a region broader than the effective diameter of the reheated clustering argument. Despite these additional complications, the ability of a solver to provide specific suggestions for improving simulation accuracy is a potentially important outcome of Data Driven computational strategies.

5.3 Data mining and multiscale Data Driven analysis

So far, we have assumed that a material data set is somehow available and provides a suitable basis for Data Driven analysis. In general, the material data may originate from a variety of sources, including experimental observations, theoretical inferences, first-principles calculations, and others. A particularly appealing possibility, which is highly synergistic with the Data Driven Computing paradigm, is the use of *multiscale analysis* to generate material data, i. e., for purposes of *Data Mining*. Data Driven Computing provides a unique foundation from which to couple the simulation scales due to its ability to directly link calculations through the determined material data associations. In effect, it lets the analysis zoom in directly on the subgrid phenomena that are causing local responses at a higher scale as visualized in Figure 5.1. Multiscale metal plasticity supplies a representative example of applications in Data Driven Computing, including the Data Driven model-free formulation of molecular dynamics, dislocation dynamics, and crystal and polycrystal plasticity.

5.3.1 From density functional theory to molecular dynamics

For metals, the foundational, or first-principles, theory is Quantum Mechanics, which characterizes the electronic structure at the subatomic level. At present, the prevailing quantum-mechanical

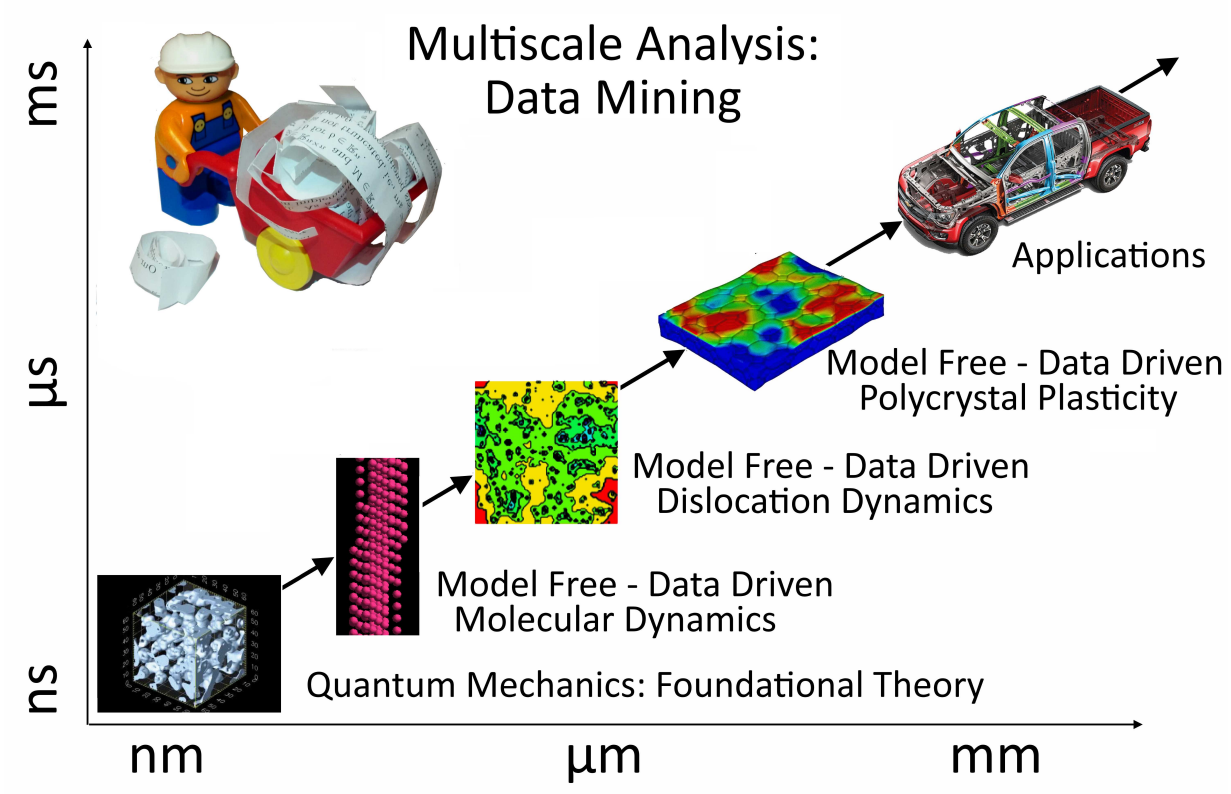


Figure 5.1: Multiscale metal plasticity as overarching application for the demonstration of Data Driven Computing. Data will be mined through multiscale analysis at all length scales. Quantum mechanics will supply the foundational theory for the multiscale hierarchy. Data Driven Computing will enable model-free calculations built on fundamental data mined from lower scales.

theory is Density-Functional Theory (DFT), which relies on models of exchange-correlation and pseudopotentials to reduce the dimensionality of the problem. Force-field data mined from DFT calculations provides a suitable basis for Data Driven Molecular-Dynamics (DDMD). The force-field space can be covered by means of *importance-sampling*, specifically by focusing on low-energy atomic configurations. Thus, whether from extensive observational evidence or from mathematical analysis [70, 28, 37, 38], it is known that such low-energy configurations consist of: i) affine deformations of a perfect lattice; ii) short-range defects of dimension zero, such as vacancies and interstitials; iii) short-range defects of dimension one, such as dislocations; and iv) short-range defects of dimension two, including free surfaces and interfaces. The aim of the DFT-based Data Mining may thus be focused on the characterization of force fields representative of said low-energy configurations in specific metals such as Al (FCC), Fe (BCC), and Mg (HCP).

5.3.2 Data Driven molecular dynamics

In using the force-field data for Data Driven Molecular Dynamics, it becomes necessary to define a distance in the phase space of atomic clusters and forces. A convenient choice of distance is supplied by 'lattice matching' [87, 88]. This distance provides a means of matching local atomic configurations against the closest configuration in the data repository and retrieving the most relevant local force field. Again we emphasize that, whereas DFT calculations have been extensively used to 'train' specific interatomic potentials [53, 67, 71, 78, 79, 92, 99], the distinguishing characteristic of Data Driven Molecular Dynamics is that the DFT data are used directly in molecular dynamics calculations, without the intermediary of an empirical interatomic potential.

5.3.3 From molecular dynamics to dislocation dynamics and plasticity

The passage from discrete-to-continuum descriptions can be effected in a number of ways, including the Quasicontinuum Method [94, 95, 62, 34], phase-field models [37, 38, 64, 63, 12, 11, 98, 29, 40, 39], and others. The resulting coarse-grained theory is Dislocation Dynamics (DD) in otherwise elastic crystals. In this coarse-grained theory, the relevant material laws are: i) the dislocation mobility law, which relates the Peach-Koehler force to dislocation velocity; ii) the core energy or line tension, which gives the core energy per unit length as a function of dislocation direction; and iii) the elastic interaction forces between dislocation segments and between dislocation segments and applied stresses. Conveniently, these laws are accessible to Data Driven Molecular Dynamics. In the case of the segment-segment elastic interaction law, the relevant phase spaces are too large to allow for uniform sampling and importance sampling must be performed instead. The interaction between distant segments can be characterized by recourse to linear elasticity [43], which restricts sampling requirements to dislocation interactions at close range. The important configurations to be sampled are low-energy configurations, specifically close-range dipoles, nodes, junctions, networks and walls [64, 44, 97, 25, 24, 30]. Finally, polycrystal plasticity data can be generated from Data Driven single-crystal plasticity by means of standard and mathematically well-understood periodic representative-volume problems [33].

5.4 Publicly-editable, open access material data repository

As already emphasized, Data Driven Computing makes use of fundamental, unbiased, model-independent data living in phase spaces set forth directly by the field equations of the problem. Thus, for instance, Data Driven Molecular Dynamics is formulated from data consisting of local atomic coordinates and corresponding forces, without reference to any particular interatomic potential and without any *a priori* assumptions on material behavior. This strict adherence to fundamental, unbiased, and model-free data not only represents a novel and significant conceptual advance but also greatly facilitates material data sharing, collection and dissemination within the community.

Specifically, *publicly-editable repository* model accessible for use by – and to contributions from – the entire scientific community become an attractive possibility. Publicly-editable repositories have demonstrated a remarkable ability for organic decentralized growth and collaborative development. Perhaps the most notable of all publicly-editable repositories is *Wikipedia* [5], the self-styled *free encyclopedia*, launched on January 15, 2001 by Jimmy Wales and Larry Sanger, which currently features over 38 million articles in more than 250 languages. As of February of 2014, *Wikipedia* had 18 billion views and 500 million unique visitors per month. *Wikipedia* forever changed the field of reference and displaced well-established and capitalized encyclopedias. What is truly remarkable about *Wikipedia*, and similar publicly-supported enterprises, is that it is free of cost, had no governmental support and grew collaboratively from anonymous contributions. *Wikipedia* was made possible by key technological advances, most notably, the internet and the development of mark-up languages, such as *Wiki*, specially designed for collaborative content modification; but it was also the result of a radically new way of thinking.

Data Driven Computing has the potential for enabling – and ushering in – a similar paradigm shift and new thinking in science and engineering, a shift towards *collaborative development of material data repositories*. Thus, by insisting on fundamental, unbiased, and model-free data, material data becomes *homogeneous* and *fungible*, and data sets of different extractions can be readily *integrated* and *fused* [14]. Data Driven Computing also makes material data repositories relevant and useful by providing the analytical and algorithmic means of integrating material data sets directly into scientific analysis or computation.

5.5 Concluding remarks

Data Driven computation represents a new direction in computational analysis that supplies new tools to the many fields of predictive science. Traditional modeling techniques have acted as a barrier between solutions and the data that provides them validity and accuracy. Removing this barrier allows for the detailing of causality flows that can exist between data sources and computational solutions. Continuing to improve the ability to link data properties and simulation outcomes represents some of the most likely extensions of the developed methods. Already work has been done to extend the work of Chapter 2 to include data interpolation for noisless data sets [45], while others have begun exploring modifications to Data Driven Computing to process visual testing outputs into extensive material data sets [86]. The efforts of this thesis and these works represent initial growths in a new data-centric simulation landscape.

Data Driven Computing offers many possible future applications for future use. The potential application of multi-scale modeling in solid mechanics discussed above represents only a first of many possible applications. More generally within mechanics, the potential reach of data-centric solution strategies touches on a common problem in a modern analysis, the need to relate response to multiple scales of analysis. These capacities arise from our new abilities of numerical science to generate data, and once such data are created, there are few restrictions on their potential use. Thus, public databases might allow for new forms of collaboration and quicker advances within the communities that study diverse arrays of phenomena. As storage costs continue to fall and computational capacity becomes ever more ubiquitous, Data Driven Computing offers data-based predictive capacity to an ever more data-rich world.

Bibliography

[1] The Knowledgebase of Interatomic Models. https://openkim.org/.

[2] The Materials Project. https://materialsproject.org/.

[3] The NIST Materials Genome Initiative. https://mgi.nist.gov/materials-data-repository/.

[4] The NoMaD Repository. http://nomad-repository.eu/cms/.

[5] Wikipedia. https://en.wikipedia.org/wiki/Wikipedia.

[6] Materials informatics. *Materials Today 8*, 10 (2005), 38 – 45.

[7] Agarwal, D., Cheah, Y. W., Fay, D., Fay, J., Guo, D., Hey, T., Humphrey, M., Jackson, K., Li, J., Poulain, C., Ryu, Y., and van Ingen, C. Data-intensive science: The terapixel and modisazure projects. *International Journal of High Performance Computing Applications 25*, 3, 304–316.

[8] Agarwal, D. A., Faybishenko, B., Freedman, V. L., Krishnan, H., Kushner, G., Lansing, C., Porter, E., Romosan, A., Shoshani, A., Wainwright, H., Weidmer, A., and Wu, K. S. A science data gateway for environmental management. *Concurrency and Computation-Practice & Experience 28*, 7, 1994–2004.

[9] Agarwal, R., and Dhar, V. Big data, data science, and analytics: The opportunity and challenge for is research. *Information Systems Research 25*, 3, 443–448.

[10] Aguilo, M. A., Swiler, L., and Urbina, A. An overview of inverse material identification within the frameworks of deterministic and stochastic parameter estimation. *International Journal for Uncertainty Quantification 3*, 4, 289–319.

[11] Argon, A. S., Xu, G., and Ortiz, M. Kinetics of dislocation emission from crack tips and the brittle to ductile transition of cleavage fracture. *Fracture-Instability Dynamics, Scaling, and Ductile/Brittle Behavior 409*, 29–44.

[12] Argon, A. S., Xu, G., and Ortiz, M. Kinetics of the crack-tip-governed brittle to ductile transitions in intrinsically brittle solids. *Cleavage Fracture*, 125–135.

[13] Arroyo, M., and Ortiz, M. Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods. *International Journal for Numerical Methods in Engineering 65*, 13, 2167–2202.

[14] Azevedo, A., and Santos, M. F. *Integration of data mining in business intelligence systems*. Business Science Reference, Hershey, 2015.

[15] Baesens, B. *Analytics in a big data world : the essential guide to data science and its applications*. Wiley & SAS business series. John Wiley & Sons, Inc., Hoboken, New Jersey, 2014.

[16] Banerjee, B., Walsh, T. F., Aquino, W., and Bonnet, M. Large scale parameter estimation problems in frequency-domain elastodynamics using an error in constitutive equation functional. *Computer Methods in Applied Mechanics and Engineering 253*, 60–72.

[17] Ben Azzouna, M., Feissel, P., and Villon, P. Robust identification of elastic properties using the modified constitutive relation error. *Computer Methods in Applied Mechanics and Engineering 295*, 196–218.

[18] Billingsley, P. *Probability and measure*, anniversary ed. Wiley series in probability and statistics. Wiley, Hoboken, N.J.

[19] Bishop, C. M. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[20] Bonnet, M., and Aquino, W. Three-dimensional transient elastodynamic inversion using the modified error in constitutive relation. *4th International Workshop on New Computational Methods for Inverse Problems (Ncmip2014) 542* (2014).

[21] Breneman, C. M., Brinson, L. C., Schadler, L. S., Natarajan, B., Krein, M., Wu, K., Morkowchuk, L., Li, Y., Deng, H., and Xu, H. Y. Stalking the materials genome: A data-driven approach to the virtual design of nanostructured polymers. *Advanced Functional Materials 23*, 46 (2013), 5746–5752.

[22] Broderick, S., and Rajan, K. Informatics derived materials databases for multifunctional properties. *Science and Technology of Advanced Materials 16*, 1.

[23] Broderick, S., Suh, C., Nowers, J., Vogel, B., Mallapragada, S., Narasimhan, B., and Rajan, K. Informatics for combinatorial materials science. *Jom 60*, 3, 56–59.

[24] Bulatov, V. V., and Cai, W. Nodal effects in dislocation mobility. *Physical Review Letters 89*, 11 (2002).

[25] Bulatov, V. V., Hsiung, L. L., Tang, M., Arsenlis, A., Bartelt, M. C., Cai, W., Florando, J. N., Hiratani, M., Rhee, M., Hommes, G., Pierce, T. G., and de la Rubia, T. D. Dislocation multi-junctions and strain hardening. *Nature 440*, 7088 (2006), 1174–1178.

[26] Ceder, G., Morgan, D., Fischer, C., Tibbetts, K., and Curtarolo, S. Data-mining-driven quantum mechanics for the prediction of structure. *Mrs Bulletin 31*, 12 (2006), 981–985.

[27] Chamoin, L., Ladeveze, P., and Waeytens, J. Goal-oriented updating of mechanical models using the adjoint framework. *Computational Mechanics 54*, 6 (2014), 1415–1430.

[28] Conti, S., Dolzmann, G., Kirchheim, B., and Muller, S. Sufficient conditions for the validity of the cauchy-born rule close to so(n). *Journal of the European Mathematical Society 8*, 3 (2006), 515–530.

[29] Conti, S., Garroni, A., and Muller, S. Singular kernels, multiscale decomposition of microstructure, and dislocation models. *Archive for Rational Mechanics and Analysis 199*, 3 (2011), 779–819.

[30] Conti, S., and Ortiz, M. Dislocation microstructures and the effective behavior of single crystals. *Archive for Rational Mechanics and Analysis 176*, 1 (2005), 103–147.

[31] Curtarolo, S., Morgan, D., Persson, K., Rodgers, J., and Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Physical Review Letters 91*, 13 (2003).

[32] Cyron, C. J., Arroyo, M., and Ortiz, M. Smooth, second order, non-negative meshfree approximants selected by maximum entropy. *International Journal for Numerical Methods in Engineering 79*, 13 (2009), 1605–1632.

[33] Dal Maso, G. *An Introduction to [gamma]-convergence*. Progress in nonlinear differential equations and their applications. Birkhauser, Boston, MA, 1993.

[34] Espanol, M. I., Kochmann, D. M., Conti, S., and Ortiz, M. A gamma-convergence analysis of the quasicontinuum method. *Multiscale Modeling & Simulation 11*, 3 (2013), 766–794.

[35] Feissel, P., and Allix, O. Modified constitutive relation error identification strategy for transient dynamics with corrupted data: The elastic case. *Computer Methods in Applied Mechanics and Engineering 196*, 13-16, 1968–1983.

[36] Feng, X., Fischer, G., Zielke, R., Svendsen, B., and Tillmann, W. Investigation of plc band nucleation in aa5754. *Materials Science and Engineering: A 539*, 0 (2012), 205 – 210.

[37] Flucher, M., Garroni, A., and Muller, S. Concentration of low energy extremals: Identification of concentration points. *Calculus of Variations and Partial Differential Equations 14*, 4 (2002), 483–516.

[38] Garroni, A., and Muller, S. Concentration phenomena for the volume functional in unbounded domains: identification of concentration points. *Journal of Functional Analysis 199*, 2 (2003), 386–410.

[39] Garroni, A., and Muller, S. Gamma-limit of a phase-field model of dislocations. *Siam Journal on Mathematical Analysis 36*, 6 (2005), 1943–1964.

[40] Garroni, A., and Muller, S. A variational model for dislocations in the line tension limit. *Archive for Rational Mechanics and Analysis 181*, 3 (2006), 535–578.

[41] Guchhait, S., and Banerjee, B. Constitutive error based material parameter estimation procedure for hyperelastic material. *Computer Methods in Applied Mechanics and Engineering 297*, 455–475.

[42] Gupta, A., Cecen, A., Goyal, S., Singh, A. K., and Kalidindi, S. R. Structure-property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system. *Acta Materialia 91* (2015), 239–254.

[43] Hirth, J. P., and Lothe, J. *Theory of dislocations*, 2nd ed. Wiley, New York, 1982.

[44] Hochrainer, T., Sandfeld, S., Zaiser, M., and Gumbsch, P. Continuum dislocation dynamics: Towards a physical theory of crystal plasticity. *Journal of the Mechanics and Physics of Solids 63* (2014), 167–178.

[45] Ibañez, R., Abisset-Chavanne, E., Aguado, J. V., Gonzalez, D., Cueto, E., and Chinesta, F. A manifold learning approach to data-driven computational elasticity and inelasticity. *Archives of Computational Methods in Engineering* (2016), 1–11.

[46] Kalidindi, S., Niezgoda, S., and Salem, A. Microstructure informatics using higher-order statistics and efficient data-mining protocols. *JOM 63*, 4 (2011).

[47] Kalidindi, S. R. Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. *International Materials Reviews 60*, 3, 150–168.

[48] Kalidindi, S. R. Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. *International Materials Reviews 60*, 3 (2015), 150–168.

[49] Kalidindi, S. R., and De Graef, M. Materials data science: Current status and future outlook. *Annual Review of Materials Research, Vol 45 45*, 171–193.

[50] Kalidindi, S. R., Gomberg, J. A., Trautt, Z. T., and Becker, C. A. Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets. *Nanotechnology 26*, 34 (2015).

[51] Kalidindi, S. R., Niezgoda, S. R., and Salem, A. A. Microstructure informatics using higher-order statistics and efficient data-mining protocols. *Jom 63*, 4 (2011), 34–41.

[52] Khinchin, A. I. *Mathematical foundations of information theory*, new dover ed. Dover Publications, New York,, 1957.

[53] Kim, S. Y., Kumar, N., Persson, P., Sofo, J., van Duin, A. C. T., and Kubicki, J. D. Development of a reaxff reactive force field for titanium dioxide/water systems. *Langmuir 29*, 25 (2013), 7838–7846.

[54] Kirchdoerfer, T., and Ortiz, M. Data-driven computational mechanics. *arXiv*, 1510.04232 (2015).

[55] Kirchdoerfer, T., and Ortiz, M. Data-driven computational mechanics. *Computer Methods in Applied Mechanics and Engineering 304* (2016), 81–101,DOI:10.1016/j.cma.2016.02.001.

[56] Kirchdoerfer, T., and Ortiz, M. Data driven computing in dynamics. *arXiv*, 1916.0954 (2017).

[57] Kirchdoerfer, T., and Ortiz, M. Data driven computing with noisy material data sets. *arXiv*, 1702.01574 (2017).

[58] Kirchdoerfer, T., and Ortiz, M. *Data-Driven Computing*. Computational Methods in Applied Sciences. Springer International Publishing, Cham, Submitted April 2017.

[59] Kirchdoerfer, T., and Ortiz, M. Data-driven computing in dynamics. *International Journal for Numerical Methods in Engineering* (Submitted June 2017).

[60] Kirchdoerfer, T., and Ortiz, M. Data driven computing with noisy material data sets. *Computer Methods in Applied Mechanics and Engineering* (Submitted March 2017).

[61] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science 220*, 4598, 671–680.

[62] Knap, J., and Ortiz, M. An analysis of the quasicontinuum method. *Journal of the Mechanics and Physics of Solids 49*, 9 (2001), 1899–1923.

[63] Koslowski, M., Cuitino, A. M., and Ortiz, M. A phase-field theory of dislocation dynamics, strain hardening and hysteresis in ductile single crystals. *Journal of the Mechanics and Physics of Solids 50*, 12 (2002), 2597–2635.

[64] Koslowski, M., and Ortiz, M. A multi-phase field model of planar dislocation networks. *Modelling and Simulation in Materials Science and Engineering 12*, 6 (2004), 1087–1097.

[65] Krein, M. P., Natarajan, B., Schadler, L. S., Brinson, L. C., Deng, H., Gai, D., Li, Y., and Breneman, C. M. Development of materials informatics tools and infrastructure to enable high throughput materials design. In *Symposium UU: Combinatorial and High-Throughput Methods in Materials Science* (2012), vol. 1425 of *MRS Online Proceedings Library*.

[66] Laarhoven, P. J. M. v., and Aarts, E. H. L. *Simulated annealing : theory and applications*. Mathematics and its applications. D. Reidel ; Sold and distributed in the U.S.A. and Canada by Kluwer Academic Publishers, Dordrecht ; Boston Norwell, MA, U.S.A., 1987.

[67] Larentzos, J. P., Rice, B. M., Byrd, E. F. C., Weingarten, N. S., and Lill, J. V. Parameterizing complex reactive force fields using multiple objective evolutionary strategies (moes). part 1: Reaxff models for cyclotrimethylene trinitramine (rdx) and 1,1-diamino-2,2-dinitroethene (fox-7). *Journal of Chemical Theory and Computation 11*, 2 (2015), 381–391.

[68] Latourte, F., Chrysochoos, A., Pagano, S., and Wattrisse, B. Elastoplastic behavior identification for heterogeneous loadings and materials. *Experimental Mechanics 48*, 4, 435–449.

[69] Liu, Z. K., Chen, L. Q., and Rajan, K. Linking length scales via materials informatics. *Jom 58*, 11 (2006), 42–50.

[70] Luckhaus, S., and Mugnai, L. On a mesoscopic many-body hamiltonian describing elastic shears and dislocations. *Continuum Mechanics and Thermodynamics 22*, 4 (2010), 251–290.

[71] Ludwig, J., Vlachos, D. G., van Duin, A. C. T., and Goddard, W. A. Dynamics of the dissociation of hydrogen on stepped platinum surfaces using the reaxff reactive force field. *Journal of Physical Chemistry B 110*, 9 (2006), 4274–4282.

[72] Merzouki, T., Nouri, H., and Roger, F. Direct identification of nonlinear damage behavior of composite materials using the constitutive equation gap method. *International Journal of Mechanical Sciences 89*, 487–499.

[73] Morgan, D., Ceder, G., and Curtarolo, S. High-throughput and data mining with ab initio methods. *Measurement Science and Technology 16*, 1 (2005), 296–301.

[74] Newman, A. R. Confidence, pedigree, and security classification for improved data fusion. *Proceedings of the Fifth International Conference on Information Fusion, Vol Ii*, 1408–1415.

[75] Nguyen, H. M., Allix, O., and Feissel, P. A robust identification strategy for rate-dependent models in dynamics. *Inverse Problems 24*, 6.

[76] Ortiz, M., Sotelino, E. D., and Nour-Omid, B. Efficiency of group implicit concurrent algorithms for transient finite element analysis. *International Journal for Numerical Methods in Engineering 28*, 12 (1989), 2761–2776.

[77] Promma, N., Raka, B., Grediac, M., Toussaint, E., Le Cam, J. B., Balandraud, X., and Hild, F. Application of the virtual fields method to mechanical characterization of elastomeric materials. *International Journal of Solids and Structures 46*, 3-4, 698–715.

[78] Psofogiannakis, G., and van Duin, A. C. T. Development of a reaxff reactive force field for si/ge/h systems and application to atomic hydrogen bombardment of si, ge, and sige (100) surfaces. *Surface Science 646* (2016), 253–260.

[79] Rahaman, O., van Duin, A. C. T., Bryantsev, V. S., Mueller, J. E., Solares, S. D., Goddard, W. A., and Doren, D. J. Development of a reaxff reactive force field for aqueous chloride and copper chloride. *Journal of Physical Chemistry A 114*, 10 (2010), 3556–3568.

[80] Rajan, K. Informatics and integrated computational materials engineering: Part ii. *Jom 61*, 1, 47–47.

[81] Rajan, K. Materials informatics. *Materials Today 8*, 10, 38–45.

[82] Rajan, K. Materials informatics part i: A diversity of issues. *Jom 60*, 3, 50–50.

[83] Rajan, K. Materials informatics: The materials "gene" and big data. *Annual Review of Materials Research, Vol 45 45*, 153–169.

[84] Rajan, K. Materials informatics how do we go about harnessing the "big data" paradigm? *Materials Today 15*, 11 (2012), 470–470.

[85] Rajan, K., Zaki, M., and Bennett, K. Informatics based design of materials. *Abstracts of Papers of the American Chemical Society 221* (2001), U464–U464.

[86] Réthoré, J. Computational measurements of stress fields from digital images. working paper or preprint, Feb. 2017.

[87] Runnels, B., Beyerlein, I. J., Conti, S., and Ortiz, M. An analytical model of interfacial energy based on a lattice-matching interatomic energy. *Journal of the Mechanics and Physics of Solids 89* (2016), 174–193.

[88] Runnels, B., Beyerlein, I. J., Conti, S., and Ortiz, M. A relaxation method for the energy and morphology of grain boundaries and interfaces. *Journal of the Mechanics and Physics of Solids 94* (2016), 388–408.

[89] Shannon, C. E. Communication theory of secrecy systems. *Bell System Technical Journal 28*, 4, 656–715.

[90] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal 27*, 3, 379–423.

[91] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal 27*, 4, 623–656.

[92] Song, W. X., and Zhao, S. J. Development of the reaxff reactive force field for aluminum-molybdenum alloy. *Journal of Materials Research 28*, 9 (2013), 1155–1164.

[93] Steinwart, I., and Christmann, A. *Support vector machines*, 1st ed. Information science and statistics. Springer, New York, 2008.

[94] Tadmor, E. B., Ortiz, M., and Phillips, R. Quasicontinuum analysis of defects in solids. *Philosophical Magazine a-Physics of Condensed Matter Structure Defects and Mechanical Properties 73*, 6 (1996), 1529–1563.

[95] Tadmor, E. B., Phillips, R., and Ortiz, M. Mixed atomistic and continuum models of deformation in solids. *Langmuir 12*, 19 (1996), 4529–4534.

[96] Warner, J. E., Diaz, M. I., Aquino, W., and Bonnet, M. Inverse material identification in coupled acoustic-structure interaction using a modified error in constitutive equation functional. *Computational Mechanics 54*, 3, 645–659.

[97] Weygand, D., Senger, J., Motz, C., Augustin, W., Heuveline, V., and Gumbsch, P. High performance computing and discrete dislocation dynamics: Plasticity of micrometer sized specimens. *High Performance Computing in Science and Engineering '08* (2009), 507–523.

[98] Xu, G., Argon, A. S., and Ortiz, M. Nucleation of dislocations from crack tips under mixed-modes of loading - implications for brittle against ductile behavior of crystals. *Philosophical Magazine a-Physics of Condensed Matter Structure Defects and Mechanical Properties 72*, 2 (1995), 415–451.

[99] Zhang, B., van Duin, A. C. T., and Johnson, J. K. Development of a reaxff reactive force field for tetrabutylphosphonium glycinate/co2 mixtures. *Journal of Physical Chemistry B 118*, 41 (2014), 12008–12016.