Computer Physics Communications 185 (2014) 2885–2892

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method

Jesper Kristensen [a], Nicholas J. Zabaras [b,*]

[a] *School of Applied and Engineering Physics, 271 Clark Hall, Cornell University, Ithaca, NY 14853-3501, USA*

[b] *School of Engineering, University of Warwick, CV4 7AL, UK*

ARTICLE INFO

Article history:
Received 13 April 2014
Received in revised form
10 July 2014
Accepted 18 July 2014
Available online 24 July 2014

Keywords:
Bayesian cluster expansion
Uncertainty quantification
Alloy modeling
High posterior density

ABSTRACT

Parametrized surrogate models are used in alloy modeling to quickly obtain otherwise expensive properties such as quantum mechanical energies, and thereafter used to optimize, or simply compute, some alloy quantity of interest, e.g., a phase transition, subject to given constraints. Once learned on a data set, the surrogate can compute alloy properties fast, but with an increased uncertainty compared to the computer code. This uncertainty propagates to the quantity of interest and in this work we seek to quantify it. Furthermore, since the alloy property is expensive to compute, we only have available a limited amount of data from which the surrogate is to be learned. Thus, limited data further increases the uncertainties in the quantity of interest, and we show how to capture this as well. We cannot, and should not, trust the surrogate before we quantify the uncertainties in the application at hand. Therefore, in this work we develop a fully Bayesian framework for quantifying the uncertainties in alloy quantities of interest, originating from replacing the expensive computer code with the fast surrogate, and from limited data. We consider a particular surrogate popular in alloy modeling, the cluster expansion, and aim to quantify how well it captures quantum mechanical energies. Our framework is applicable to other surrogates and alloy properties.

1. Introduction

In the present work, we aim to develop a Bayesian framework for quantifying the uncertainty in alloy modeling when using fast parametrized surrogates in place of an expensive computer code. In the most typical setup, the surrogate is learned from some data set, e.g., quantum mechanical energies, and then used to predict some *quantity of interest* (QI), which could be a ground state line, a phase transition, or some optimal structure (e.g., lowest thermal conductivity in the case where the data are instead thermal conductivities). Of course, the parametrization we choose for the surrogate depends on what data we obtain from the computer code. Since the code is expensive, we only have available a limited amount of data. Furthermore, we require the surrogates to be computationally cheap. This means that, e.g., if the surrogate is represented by a set of basis functions, we are not in liberty to include an arbitrarily large number of such basis functions. The particular surrogate we consider later is such an example. These restrictions on the surrogate mean that, when it is parametrized, we do not know *a priori* the best parametrization. We have to learn it from a set of multiple candidate parametrizations, a pool of candidates, each candidate, from a Bayesian perspective, consistent with the observed limited amount of data. A single value of the QI is computed from a single surrogate candidate. Since there may be multiple candidates, there may also be multiple values for the same QI. Our uncertainty about the best surrogate candidate has thus propagated to the QI. This is the first source of uncertainty we aim to capture in the present work. From now on, we will simply say parametrization to mean surrogate parametrization/candidate.

Notice also that the effect of limited data enters implicitly through our belief about the best parametrization pool to choose. For example, upon seeing data set $\mathcal{D}_1$ it might be that the pool of parametrizations t_1 is better than another pool t_2. But if we now observe more data, it could very well be that our opinion is reversed, thus choosing t_2 over t_1. The fewer data points we have, the worse, and generally larger, our pool of parametrizations consistent with the data will be, unless, of course, our prior belief is already sharply tuned to a good pool of parametrizations. However, this is rarely the case, and in by far the most cases we benefit from observing data. From this, it should be clear that the limited data plays a role in our knowledge about the best pool of parametrizations to use.

* Corresponding author.
E-mail addresses: nzabaras@gmail.com, zabaras@cornell.edu (N.J. Zabaras).
URL: http://www.zabaras.com (N.J. Zabaras).

http://dx.doi.org/10.1016/j.cpc.2014.07.013
0010-4655/© 2014 Elsevier B.V. All rights reserved.

The fact that we only see a limited amount of data therefore introduces a second source of uncertainty (not independent of the first one though) in the QI, and we will be able to capture this as well.

Our developed methods are independent of the particular surrogate employed, but we will focus on a very popular choice in materials science: the cluster expansion [1]. The cluster expansion expands the alloy property in basis functions with associated expansion coefficients called *effective cluster interactions* (ECI) [2]. It is useful in capturing properties, which depend on the particular atomic arrangement on the lattice, this arrangement being called a configuration. It has been used to describe quantum mechanical energies, thermal conductivities, band gaps [3], etc., of a multitude of alloys. The ECI are obtained by fitting the cluster expansion to a data set. The cluster expansion surrogate is uniquely given once the ECI are specified, so we will consider a surrogate parametrization as being synonymous with ECI. Although the cluster expansion is exact when untruncated, in practice one needs to make a truncation choice and estimate the ECI from a pool of parametrizations, as discussed earlier. We reiterate that this introduces uncertainties in the QI predicted by the cluster expansion. Although an important question to ask, the sizes of these uncertainties have remained unknown until now. We employ a fully Bayesian approach to quantify these uncertainties.

We should mention that non-Bayesian methods have been applied in some works to quantify the uncertainty in QI's, but we believe that they should be avoided for uncertainty propagation. In particular, there is no rigorous framework for propagating uncertainties through parametrized surrogates, such as the cluster expansion, to the QI in non-Bayesian frameworks [4,5].

In beginning our Bayesian approach, we need to be clear about what we mean by *probability*. We interpret probability as a reasonable degree of belief [6,7] as opposed to a frequency of some (hypothetical) long-run experiment. The sum and product rules of probability theory then tell us how to manipulate degrees of belief in a rigorous way. From this view of probability, the Bayes theorem follows and can be used to change our knowledge when observing new data in a given problem [8,6]. This is collectively what is called Bayesian probability theory. We will use a Bayesian approach to introduce a model describing our belief about the best set of ECI with emphasis on sparsity. We include the sparsity feature because alloy properties are expected to be sparsely representable, based on physical arguments [9]. A very successful sparse regression method, from the non-Bayesian literature, is the least absolute shrinkage and selection operator method (LASSO), which is an L_1-constrained least squares method [10]. It can be shown that LASSO has a Bayesian interpretation. It corresponds to the posterior mode when the parameters to be learned have independent Laplace distributions as priors [11]. The above information about LASSO will be used to choose Laplace distributed priors in Section 2.4. The Bayesian posterior distribution (posterior) contains the information needed to rigorously quantify the QI uncertainties. In our case, the posterior attains a shape allowing it to be summarized via the 95% highest posterior density confidence interval (HPD)—the smallest region containing at least 95% of the posterior mass. We will reduce the effects of other uncertainties as much as possible and discuss this as we go along.

We employ our framework to two real binary alloy systems. First, we consider body-centered-cubic (bcc) magnesium–lithium (Mg–Li) and let the QI be its ground state line. Then, we turn to diamond silicon–germanium (Si–Ge) and present a computationally more involved example where the QI is the transition temperature of the disordered to two-phase-coexistence at 50% composition.

The paper is organized as follows. We start out with a general introduction to uncertainty quantification and present our framework in Sections 2.1 and 2.2. In Section 2.3 we discuss where and how the cluster expansion enters the scene. Then, in Section 2.4, we present a Bayesian method for describing the ECI with emphasis on sparsity. This posterior will not be in closed form for its intended use so we show how samples are drawn from it using MCMC methods in Section 2.5. Having developed the framework we turn to case studies first discussing uncertainty quantification in the Mg–Li ground state line in Section 2.6, followed by the uncertainty quantification of an Si–Ge phase transition in Section 2.7. Results from carrying out these are presented in Section 4 and a corresponding discussion follows in Section 5. The paper is concluded in Section 6.

2. Uncertainty quantification

2.1. Background

In this section we introduce the methods used to quantify the uncertainty in the QI making no assumption about the form of parametrization of the response surface. Then, in the following section, we show how the cluster expansion makes this parametrization. Independent of the choice of surrogate model we will need data to make the best possible choice of parametrization. Therefore, we first discuss assumptions about the computer code used to obtain the data. Then, we introduce the central element in this work: an operator which acts on the surrogate to produce the QI, and show how it is used to summarize the present uncertainty quantification task in a single equation, given certain assumptions.

The data acquisition takes place by supplying a set of alloy configurations as input to an expensive computer code, e.g., VASP [12–18] or LAMMPS [19], and obtain a set of corresponding property values as output which collectively form the response surface. This could be quantum mechanical energies per atom or thermal conductivity, respectively. We view the computer code as a function $f(\cdot)$ mapping some input structure, with configuration denoted as $\boldsymbol{\sigma}$, to a response y. We do not know $f(\cdot)$ and we are most often not interested in it *per se*, but rather some function of this—the QI. Therefore, we define an operator $I[\cdot]$ taking as input a response surface and returning the QI which can generally be represented by a set of real numbers. As an example, we can let it return the structure at the global minimum of the surface:

$$I[f(\cdot)] = arg\,\min_i f(\boldsymbol{\sigma}_i),$$

or the convex hull of formation energies:

$$I[f(\cdot)] \\ = \left\{ \sum_j \lambda_j \Delta f(\boldsymbol{\sigma}_j) : \lambda_j \geq 0 \text{ for all } j \text{ and } \sum_j \lambda_j = 1 \right\}, \tag{1}$$

where $\Delta f(\boldsymbol{\sigma}_j)$ is the formation energy of structure j (dependent on $f(\boldsymbol{\sigma}_j)$) defined later in Eq. (13) but with $E(\boldsymbol{\sigma}_j)$ replacing $f(\boldsymbol{\sigma}_j)$. Another example, related to the convex hull, is the ground state line. The QI can also be more complicated such as a transition temperature. The computer code has inherent approximations. For example, VASP approximates the exchange–correlation term in density functional theory (DFT), implements a particular k-point integration scheme [20], and a pseudopotential to approximate the true potential [21], etc., all inducing uncertainties in the output. We assume, however, that such uncertainties are small when compared to those arising from not knowing how to choose the surrogate parametrization, and from having observed only a limited amount of data. The presence of code uncertainties means that we do not actually observe the theoretical $f(\boldsymbol{\sigma}_i)$ of structure i. Nevertheless, we will make the assumption that we do observe $f(\cdot)$, but with added Gaussian measurement noise. Call this noisy version of the code y_i. We show in Section 2.4 how the noise can be estimated in the current framework. Incidentally, this does not mean we commit to the noise source necessarily being Gaussian itself, but rather