A Bayesian approach to evaluating the uncertainty of thermodynamic data and phase diagrams

M. Stan*, B.J. Reardon

Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received 16 September 2003; accepted 26 November 2003

Abstract

A heuristic optimization methodology based on Bayesian statistics is presented. The goal is to help researchers decide on the optimal set of thermodynamic data to use. This approach accounts for the errors associated with reported data and how reliable the researcher believes the model to be. The optimization is conducted with a multi-objective genetic algorithm (GA) coupled with Bayesian statistics to more accurately link the limited and uncertain experimental thermodynamic data to thermodynamic models of interest. The computer program provides guidance as to which experiments are needed to enhance the reliability of the dataset and is ideally suited for parameter optimization and sensitivity analysis. Applications include the UO₂–PuO₂ and UO₂–BeO systems.

© 2004 Elsevier Ltd. All rights reserved.

1. Introduction

Finding an optimal model by fitting thermodynamic data is a difficult problem in materials science due to the high uncertainty associated with the experimental or calculated data sets that are used as input. This situation is prevalent in the case of the phase diagram calculations [1] where the solidus and liquidus boundaries are uncertain [2]. In spite of that, one could rarely find the uncertainty limits reported together with the phase diagram.

One could address this problem in several ways, each having a number of limitations. The first would be to simply accept a particular set of thermodynamic data as fact and use these values to calculate phase equilibrium curves. This approach ignores all the other data available in other thermodynamic assessments as well as any phase diagram data. It also assumes that the selected data set is the best available. The second approach would be to take an average of all the known thermodynamic data sets. This assumes that all the thermodynamic data is equal in quality and thus only an average is necessary. This approach also ignores the available phase diagram data. Another option is to fit the solidus and liquidus equations to the known phase diagram data. The starting point of the optimization heavily influences this approach. Furthermore, it essentially ignores the experimental thermodynamic data once the optimization has initiated. The method proposed in this work uses a genetic algorithm with Bayesian statistics to incorporate all the data and its associated uncertainties into an optimal fit of what is known.

The scientific literature is notably bereft of papers dedicated to the analysis of the uncertainties associated with equilibrium phase diagrams. A classical approach, based on the “spread of mistakes” formalism, is presented in [3] while in [4] a way of extracting the maximum information from a minimal set of experimental data is investigated. Bayesian based methods have been used to produce self-consistent thermodynamic data sets for binary [5] or multi-component [6–8] systems. All papers emphasize the importance of starting with a reliable, self-consistent thermodynamic data set and draw the reader’s attention to the propagation of the errors in the input parameters during the calculation process.

This work demonstrates that modern heuristic optimization techniques such as genetic algorithms combined with Bayesian statistics offer a viable way of relating models to the data in the face of large uncertainties both on the model parameters and the training data. To illustrate the utility of this approach the solidus and liquidus boundaries of the PuO₂–UO₂ and BeO–UO₂ systems have been selected. The parameter optimization was performed for each binary system A–B given some information about the solidus and
liquidus boundaries, the heats of melting $\Delta H^M$ and the melting temperatures $T^M$.

2. Bayesian statistics

Bayesian statistics, first proposed by Rev. Thomas Bayes [9] and discussed in many books since (see Antelman [10] as an example) provides a formal way of determining the degree of belief in a model given available data using the law of conditional probabilities. To illustrate this approach in a binary system $A$–$B$, consider a model parameter vector that must be optimized:

$$M = (\Delta H^M_A, \Delta H^M_B, T^M_A, T^M_B)^T$$

where $\Delta H^M$ and $\Delta T^M$ are enthalpy and temperature of melting, respectively, and $V^T$ is the transpose of the vector $V$.

A data vector is defined as:

$$D = (D_1, D_2, \ldots, D_N)^T$$

where $N$ is the number of data points.

The goal of Bayesian analysis is to come up with a way of accepting or rejecting a particular model $(M)$ or hypothesis given an experimental data set $(D)$ and prior knowledge about the problem. Thus, in Bayesian statistics, the model or hypothesis is assigned a probability of acceptance and the total probability distribution function (PDF) of a series of models being tested makes up what is commonly called the posterior probability density (PPD). This goal is achievable through the central tenet of Bayesian statistics, Bayes’s theorem:

$$\sigma(M_j \mid D) = \frac{P(D \mid M_j)P(M_j)}{\sum_i P(D \mid M_i)P(M_i)} \quad (1)$$

where $i$ is a summation index over the collection of $n$ models. $P(M_j)$ is the prior probability for the model $M_j$ to be correct, which is often based on subjective knowledge about the system. $P(D \mid M_j)$ is the likelihood function, showing how well the data set $D$ is described by the model $M_j$, and $\sigma(M_j \mid D)$ is the posterior probability for the model $M_j$ to be correct, given the new data set $D$.

The theorem states that the PPD $(\sigma(M \mid D))$ or the conditional probability of the model being correct given the data is a ratio of the PDF of $M$ and $D$ to the PDF of $D$ alone. Once $\sigma(M \mid D)$ was determined, several useful statistical results are obtained, such as the mean model:

$$\langle M \rangle = \sum_i M_i \sum_j \sigma(M_i \mid D_j); \quad (2)$$

the posterior covariance matrix:

$$C = \sum_i M_i M_i^T \sigma(M_i \mid D) - \langle M \rangle \langle M \rangle^T; \quad (3)$$

3. Genetic algorithm

Genetic algorithms are heuristic optimization techniques that borrow heavily from the ideas of Darwinian evolution [12]. Using evolution as an optimization tool was first proposed by Holland in 1975 [13] and ever since has spurred a large amount of interest [14]. A genetic algorithm borrows the three main constructs of Darwinian evolution (selection, crossover, and mutation) to evolve a set of parameter vectors towards an optimal solution.

The key to the successful implementation of a GA for this problem lies in the selection operator. The multi-objective optimization literature describes many ways to perform selection in such a world. The most common incorporates the concept of the Pareto optimal frontier [15]. One member is said to dominate another if it is at least as good as it is in all objectives and strictly better in one objective. The concept of Pareto optimality allows one to rank the members of a population according to their degree of dominance. The members that are not dominated by any other members are said to be rank one and thus the most fit. The members that are only dominated by the rank one members are said to be of rank two and not as fit as the rank ones. The concept of Pareto optimality generally works well for small numbers of objectives (2 to 3). A problem arises as the number of objectives increases since the likelihood of dominance decreases [16, 17]. Thus, as the number of objectives increases there is an increasing likelihood that all the members become rank one and thus there is no clear way of selecting one member over the other. This problem is compounded by the fact that the objective functions of this problem must use experimental data that have a large degree of uncertainty. Uncertainty in the experimental data relaxes the constraints on the optimal parameter values and...
forces us to identify a range of parameter values that provide a range of calculated values that lie within the experimental uncertainty. Thus, this problem is not only multi-objective but it is also under specified. Previous work [11] has shown that selection using a fuzzy logic-weighting scheme can be more useful than Pareto optimality.

A fuzzy logic-weighting scheme [18] looks at all the objective values of a particular member and rescales them to a value between 0 and 1. Zero if the value is the worst of the population. One if it falls within the experimental uncertainty. Once the objectives have been scaled, the average is taken over all objectives and that single number is the fitness for the member in the population.

Using the fuzzy logic-weighting scheme, the GA is run until all the members of the population reach a fitness of 1 or at least reach a state of equilibrium where there is no more improvement. When this state is reached, the members of the final population are used to determine the uncertainty bounds on the model parameters. The population of final parameters can then be used to bound output of the model and show where the model is most uncertain and in need of more data.

There are a number of advantages to using a GA in this problem over other fitting routines. First, a multi-objective GA tends to be robust enough to identify distributions of solutions. These distributions are often multi-modal and thus have shapes not easily captured by traditional fitting routines. Second, and probably most importantly, the GA incorporates all the known data into its search. For example, the known phase diagram data defines the objectives and the spread of the known thermodynamic data defines the search space. As with any stochastic process, small variations in the solution parameters occur in different runs. The user must define the acceptance criteria for the solution.

### 4. Results and discussion

The UO$_2$–PuO$_2$ system shows complete solubility of the two components in the solid phase [19]. The liquidus ($x^\text{Liq}$) and solidus ($x^\text{Sol}$) mole fractions for each fixed temperature ($T$) can be approximated [20, 21] by:

$$x^\text{Liq}(T) = \frac{1 - \exp \left( \frac{\Delta H^M_{\text{UO}_2}}{R} \left( \frac{1}{T - T^\text{M}_{\text{UO}_2}} \right) \right)}{\exp \left( \frac{\Delta H^M_{\text{UO}_2}}{R} \left( \frac{1}{T - T^\text{M}_{\text{UO}_2}} \right) \right) - \exp \left( \frac{\Delta H^M_{\text{PuO}_2}}{R} \left( \frac{1}{T - T^\text{M}_{\text{PuO}_2}} \right) \right)}$$

(7)

and

$$x^\text{Sol}(T) = x^\text{Liq}(T) \cdot \exp \left( \frac{\Delta H^M_{\text{PuO}_2}}{R} \left( \frac{1}{T - T^\text{M}_{\text{PuO}_2}} \right) \right)$$

(8)

where $R$ is the gas constant (8.314 J/mol K).

The values of the input parameters for the UO$_2$–PuO$_2$ system are scattered [22–27], leading to large uncertainty limits, as shown in Table 1. The goal of our work is to further refine the values of the input parameters using the GA and Bayesian statistics given known experimental data on the solidus and liquidus positions (Fig. 1). Thus the optimization proceeds as follows: first, the search range for each parameter is defined for the GA based on the accepted uncertainty or variation in the published parameter values. Second, the GA evolves the parameter values based on how well the values generate solidus and liquidus curves, which match the available experimental data—taking into account the uncertainties of the experimental solidus and liquidus curves. For this study the initial uncertainty in concentration was assumed to be 0.05 and the uncertainty in the liquidus and solidus temperatures was 55 K and 35 K respectively. Once the range of parameter values is optimized, values are sampled from this range to determine

![Fig. 1. The calculated UO$_2$–PuO$_2$ phase diagram (solid lines) and the uncertainty intervals of the liquidus (dashed lines) and the solidus (dotted lines) curves.](image-url)
a range of values in the forward calculated solidus and liquidus curves. This results in fuzzy or uncertain bands that define the most probable position of the curves given all data and the underlying model. These bands are displayed using error bars in Fig. 1. The optimized values of the melting enthalpy and temperature are displayed in Table 2. Of particular importance to note is that the total uncertainty in model parameter values as well as the experimental phase diagram has decreased by using this method of heuristic optimization and Bayesian statistics.

To optimize the UO₂–PuO₂ system, a similar procedure as for the UO₂–PuO₂ system was employed. For this type of diagram the equilibrium lines are defined by:

\[ x_{\text{UO}_2+\text{Liq}}(T) = 1 - \exp \left( \frac{-\Delta H_{\text{UO}_2}^M}{RT} \ln \left( \frac{T_{\text{M}_{\text{UO}_2}}}{T} \right) \right) \]  

(9)

and

\[ x_{\text{Liq}+\text{BeO}}(T) = \exp \left( \frac{-\Delta H_{\text{BeO}}^M}{RT} \ln \left( \frac{T_{\text{M}_{\text{BeO}}}}{T} \right) \right) \]  

(10)

The values of the melting enthalpy and temperature were obtained from the literature and are displayed in Table 1. Note that the values for UO₂ differ from the previous study. The UO₂–BeO phase diagram [28] shows a eutectic point at \( T = 2450 \) K and BeO mole fraction \( x = 0.68 \). Namely \( x_{\text{BeO}} = 0.68 \pm 0.05 \) (Fig. 2). For this study the uncertainty in the liquidus concentration was again 0.05 and the temperature uncertainty was 40 K. In the model, the eutectic composition is defined as that point in which the curves calculated from Eqs. (9) and (10) cross. This point also defines the calculated eutectic temperature. While the eutectic temperature is known experimentally, there is no information gain in comparing it to the calculated value since the calculated value is determined by the calculated value of the eutectic composition.

As in the previous study, the rank one Pareto members are collected and statistics on the optimized parameter and objective values are calculated. The net result is an average and standard deviation of the variable values as shown in Table 2. These variable combinations are inserted into the forward objective equations to obtain calculated distributions on the solidus and liquidus boundaries as shown in Fig. 2. The temperature and composition uncertainty associated with the eutectic point was derived from the intersections of the uncertainty bands of the two liquidus lines. The net result is an apparent fuzziness of the calculated phase boundaries. This fuzziness is ultimately due to the uncertainty and sparseness of the experimental data.

The extension of this method to multi-component systems is immediate when all phases are ideal solutions or stoichiometric compounds. However, for complex phase diagrams, analytical functions are not available to describe the equilibrium lines. Instead, the temperature can be calculated at each composition point by solving for the equality of the chemical potentials of each component in all phases that are present in the equilibrium configuration. This method can be further developed to incorporate an iterative procedure for such calculation or can become a module in existing phase diagram calculations software.

5. Conclusions

The use of a genetic algorithm with Bayesian statistics allows for incorporating uncertain data sets, both large and small, in an efficient and meaningful way. This process then leads to the optimization of the parameters of proposed models and the assessment of the overall predictive credibility of said models. Specifically, this work determines the degree of uncertainty on the phase boundaries of the UO₂–PuO₂ and UO₂–BeO systems by taking into account the available phase boundary data, the accepted models of the phase boundaries, and the thermodynamic data used in those models. The net result was an overall reduction in uncertainty of the values of the thermodynamic data as well
as the phase boundary positions. The use of modern heuristic optimizers such as genetic algorithms was crucial to this work since they are both robust and require no assumptions about the forms of the uncertainty distributions.

References