2nd World Congress on Integrated Computational Materials Engineering
Edited by: Mei Li, Carelyn Campbell, Katsuyo Thornton, Elizabeth Holm, and Peter Gumbsch
TMS (The Minerals, Metals & Materials Society), 2013

CROSS-SCALE, CROSS-DOMAIN MODEL VALIDATION BASED ON GENERALIZED HIDDEN MARKOV MODEL AND GENERALIZED INTERVAL BAYES' RULE

Yan Wang[1], David L. McDowell[1], Aaron E. Tallman[1]

[1]Georgia Institute of Technology
801 Ferst Drive NW, Atlanta, GA 20332-0405

Keywords: Irradiation, Uncertainty Quantification, Imprecise Probability, Hidden Markov Model

Abstract

Reliable simulation protocols supporting integrated computational materials engineering (ICME) requires uncertainty to be quantified. In general, two types of uncertainties are recognized. Aleatory uncertainty is inherent randomness, whereas epistemic uncertainty is due to lack of knowledge. Aleatory and epistemic uncertainties need to be differentiated in validating multiscale models, where measurement data for unconventionally very small or large systems are scarce, or vary greatly in forms and quality (i.e., sources of epistemic uncertainty). In this paper, a recently proposed generalized hidden Markov model (GHMM) is used for cross-scale and cross-domain information fusion under the two types of uncertainties. The dependency relationships among the observable and hidden state variables at multiple scales and physical domains are captured by generalized interval probability. The update of imprecise credence and model validation are based on a generalized interval Bayes' rule (GIBR).

Introduction

Not all uncertainties or errors in simulation can be readily represented by classical probability distributions. In general, two types of uncertainties are recognized. Aleatory uncertainty is inherent randomness, whereas epistemic uncertainty is due to lack of knowledge. Epistemic uncertainty is a result of conflicting information from multiple sources, conflicting beliefs among experts' opinions, lack of data, lack of time for introspection, measurement error, lack of dependency information, truncation errors during numerical treatments, etc. In contrast to aleatory uncertainty, epistemic uncertainty can be reduced.

In modeling and simulation (M&S), epistemic uncertainty manifests from errors associated with the models and input data. For instance, in finite-element analysis, epistemic uncertainty arises from the truncation error involved in linear and nonlinear approximations of strain fields using polynomials, imprecise input parameters in the model, mismatch of structures and geometric configurations between the modeled and the true physical ones, etc. In kinetic Monte Carlo simulation, epistemic uncertainty is largely from the imperfect knowledge and resulting simplification of transition paths and reaction rates, as well as dynamic rates because of external loads, crowding effects, and other unknown correlations. In molecular dynamics simulation, epistemic uncertainty is mainly from the inaccurate potential functions (both forms and parameters), boundary conditions, local approximation of potentials, cut-off distance, high strain rates/short simulation time scales, as well as measurement errors in experimental data during model calibration. In contrast to epistemic uncertainty, various sources such as disturbed boundary and loading conditions, different sampling sizes and running times, inherent variations

of material compositions, and other randomness and fluctuation contribute to the aleatory component of uncertainty.

Since aleatory and epistemic uncertainties arise from distinct sources and have very different characteristics, we should represent them differently, preferably in different forms, so that they can be easily differentiated. Aleatory uncertainty is traditionally and predominantly modeled using probability distributions. In contrast, epistemic uncertainty has been modeled in several ways, such as probability, interval, fuzzy set, random set, basic probability assignment, etc. Here we use intervals to quantify epistemic uncertainty.

Interval is as simple as a pair of numbers, i.e. the lower and upper bounds. The reason to choose an interval representation is two-fold. First, an interval is a natural means for human users to communicate information and is simple to use. It has been widely used to represent a range of possible values, an estimate of lower and upper bounds for numerical errors, and the measurement error because of the available precision as the result of instrument calibration. Second, an interval can be regarded as the most suitable way to represent the lack of knowledge. Compared to other forms, specification via an interval has the least assumptions. It only needs lower and upper bounds, without any assumption of distributions between them. Given that epistemic uncertainty arises intrinsically from lack of knowledge, a representation with the least assumptions is most desirable.

Probability has certain limitations in representing epistemic uncertainty. The accuracy of a predictive simulation depends heavily on fundamental understanding of the underlying physical and chemical processes. Lack of perfect knowledge and fundamental insight inevitably renders models imperfect. Any assumption regarding distributions in M&S introduces a bias. The most significant limitation of probability is that it does not differentiate 'total ignorance' from other probability distributions. A problem arises because introducing a uniform or any particular form of distribution implicitly introduces extra information that cannot be justified in the case of zero knowledge. This leads to the Bertrand-style paradoxes. "Knowing the unknown" does not represent total ignorance. Although the Bayesian approach has been proposed to reduce the bias introduced in assuming a distribution, and it serves the purpose well in an ideal situation where we have plentiful data without measurement errors, its limitation remains in the real-world applications where lack of data or imperfect measurement lingers. Therefore, it is desirable to have more general and efficient approaches to incorporate epistemic uncertainty in M&S, with minimal assumptions regarding probability distributions and their parameters. Moreover, it is desired that such approaches be less computationally demanding than the traditional Bayesian learning and update approach.

Basic Elements of Generalized Interval Probability

Interval or imprecise probability, represented by lower and upper probability bounds as $[\underline{p}, \overline{p}]$, is a generalization of classical probability that simultaneously models the two uncertainty components; aleatory uncertainty is modeled by probability whereas epistemic uncertainty by interval. When $\underline{p} = \overline{p}$, the degenerated interval probability becomes the traditional probability. Differing from other forms of imprecise probabilities, such as Dempster-Shafer theory [1,2], coherent lower prevision [3], p-box [4], etc., the recently proposed generalized interval probability [5] provides a simplified probabilistic calculus structure that ensures ease of application.

Generalized Interval Probability

A generalized interval $\mathbf{x} \coloneqq [\underline{x}, \overline{x}]$ is defined as a pair of numbers. $[\underline{x}, \overline{x}]$ is called *proper* if $\underline{x} \le \overline{x}$, and *improper* if $\underline{x} \ge \overline{x}$. The introduction of improper intervals greatly simplifies the calculus structure of interval probability and makes it very similar to the one in the classical precise probability theory. The calculation of generalized intervals is based on the Kaucher interval arithmetic [6]. In generalized interval probability theory, the probability measure has the value of generalized interval. Therefore, both [0.2,0.4] and [0.4,0.2] are valid probabilities. The relationship between proper and improper intervals is established by a *dual* operator. For instance, dual[0.2,0.4]=[0.4,0.2] and dual[0.4,0.2]=[0.2,0.4].

In generalized interval probability, *conditional probability* is uniquely defined as $\mathbf{p}(X \mid Y) \coloneqq \mathbf{p}(XY) / \operatorname{dual} \mathbf{p}(Y) = [\underline{p}(XY) / \underline{p}(Y), \overline{p}(XY) / \overline{p}(Y)]$. As a result, the *generalized interval Bayes' rule* states that $\mathbf{p}(E_i \mid A) = \mathbf{p}(A \mid E_i)\mathbf{p}(E_i) \Big/ \sum_{j=1}^{n} \operatorname{dual} \mathbf{p}(A \mid E_j) \operatorname{dual} \mathbf{p}(E_j)$ where E_i (i=1,..,n) are mutually disjoint events as the partition of the sample space, and most importantly $\sum_{j=1}^{n} \mathbf{p}(E_j) = 1$, which is called the logic coherent constraint (LCC). With simple algebraic calculation, the probabilistic reasoning based on generalized interval probability is very similar to the traditional one in the classical probability. In contrast, other forms of imprecise probabilities must rely on linear or nonlinear optimization methods to estimate probability lower and upper bounds, which is computationally cumbersome.

Generalized Hidden Markov Model (GHMM)

A GHMM [7] was recently proposed to capture correlations of variables between scales, as illustrated in Fig. 1, where dependencies between random variables at three length scales Ω_x, Ω_y, and Ω_z are captured. The state variables, denoted as x_i, y_j, and z_k, respectively associated with these three scales, are hidden. Their true values have to be predicted and inferred by some measurable or observable quantities X_i, Y_j, and Z_k, respectively, via physical experiments. The correlation relationships are expressed as conditional probabilities. For instance, $\mathbf{p}(x_i|x_{i,1}, x_{i,2})$ captures the dependency within one scale, $\mathbf{p}(y_j|x_{i,1}, x_{i,2})$ captures the one between scales, and $\mathbf{p}(X_i|x_i)$, $\mathbf{p}(Y_j|y_j)$, and $\mathbf{p}(Z_k|z_k)$ are between the observable and hidden variables. It should be noted that these 'scales' may also be constituted by domains (e.g., time, different loosely coupled model environments), so that the conditional probabilities can be expressed not only in terms of disparate length and time scales, but associated sets of models, even multiple models.

Application of GHMM in Model Validation under Uncertainties

Similar to the Bayesian approach in model validation [8, 9], the validation of simulation models under both epistemic and aleatory uncertainties at multiple scales can be performed based on the generalized interval Bayes' rule. Given the experimental measurements Y_1, …, Y_M and Z_1, …, Z_N in separate domains or scales, the interval probability that the model has the parameter x is

$$\mathbf{p}(x \mid Y_1,\ldots,Y_M,Z_1,\ldots,Z_N) = \frac{\mathbf{p}(x)\int\cdots\int\begin{bmatrix}\mathbf{p}(Z_1,\ldots,Z_N \mid z_1,\ldots,z_N)\mathbf{p}(Y_1,\ldots,Y_M \mid y_1,\ldots,y_M)\\ \mathbf{p}(z_1,\ldots,z_N \mid y_1,\ldots,y_M)\mathbf{p}(y_1,\ldots,y_M \mid x)\end{bmatrix} dz_1\cdots dz_N dy_1\cdots dy_M}{\operatorname{dual}\int\cdots\int\begin{bmatrix}\mathbf{p}(Z_1,\ldots,Z_N \mid z_1,\ldots,z_N)\mathbf{p}(Y_1,\ldots,Y_M \mid y_1,\ldots,y_M)\\ \mathbf{p}(z_1,\ldots,z_N \mid y_1,\ldots,y_M)\mathbf{p}(y_1,\ldots,y_M \mid x_1,\ldots,x_L)\mathbf{p}(x)\end{bmatrix} dz_1\cdots dz_N dy_1\cdots dy_M dx} \quad (1)$$

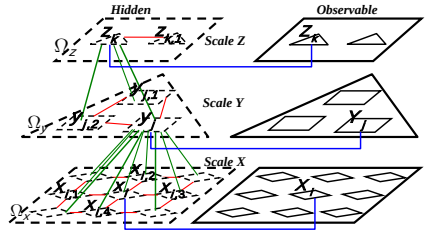

Figure 1. The generalized hidden Markov model captures spatial and scale dependencies.

where $y_1, \ldots, y_M$ and $z_1, \ldots, z_N$ are hidden variables of physical properties at different domains or scales and related to x. In the numerator of Eq.(1), $\mathbf{p}(z_1,\ldots,z_N|y_1,\ldots,y_M)$ and $\mathbf{p}(y_1,\ldots,y_M|x)$ represent the correlations between hidden variables of properties at different scales, whereas $\mathbf{p}(Z_1,\ldots,Z_N|z_1,\ldots,z_N)$ and $\mathbf{p}(Y_1,\ldots,Y_M|y_1,\ldots,y_M)$ capture the correlations between hidden and measurable properties. The key to applying Eq. (1) is to find conditional probability values from experiments and prior experience.

An Example of A Molecular Dynamic Simulation Model for Irradiation on Fe

Here we demonstrate how to validate models using the generic GHMM approach with a molecular dynamics (MD) simulation of point defect generation in Fe crystals that are subject to high energy knock-on atom collisions (irradiation) and resulting cascade events. A MD simulation model of the Fe crystal was constructed in LAMMPS. For each combination of different energy levels and radiation angles, 16 simulation runs are conducted. The probabilities that a stable Frenkel pair are generated, also equivalently known as damage function $\nu(T)$ with the transfer or recoil energy T, are collected and shown in Fig. 2 with the label 'mid'. Because of the model and numerical errors involved, the probabilities do not necessarily always increase as the energy level increases, as shown in Fig. 2(b) and (c). Therefore, interval probabilities with lower and upper limits are used. The interval probability widths are calculated by the standard deviation of binomial distributions, where error bounds are added to and subtracted from the middle or nominal values of probabilities. Interval probabilities capture the uncertainties and errors involved in the MD simulation model. To ensure the non-decreasing pattern of the lower and upper cumulative distribution functions (CDF) $P(T \le t)$, a filtering procedure is taken to 'smooth out' the empirical CDF's as follows. If the upper CDF value at a given energy level is less than the one at a lower energy level (immediately on its left in the chart), its value is set to be the same as the one at the immediate lower energy level. Similarly, if the lower CDF value at an energy level is larger than the one at a higher energy level (immediately on its right in the chart), its value is set to be the same as the one at the immediate higher energy level.

During physical experiments, the total displacement cross section σ, as the indicator of the amount of point defects, is not measured directly. Rather, it is based on the measurement of the electrical resistivity change rates $\Delta\rho/n$ that is correlated with σ. Instead of using the traditional empirical analytical relationship between $\Delta\rho/n$ and σ where uncertainty is ignored, we consider the correlation between the two in terms of conditional probability. Some examples of the correspondence between the two quantities from experimental measurement are shown in Table I [10], where the measurement error is listed in terms of the minimum and maximum values of $\Delta\rho/n$. In other words, at a particular level of maximum transmitted energy T_m, one value of σ

corresponds to a range of possible values of $\Delta\rho/n$ and a uniform distribution between the bounds can be assumed. The correlation between the two quantities thus can be represented as conditional probability $P(\Delta\rho/n|\sigma,T_m)$. Similarly, the correspondence between the total displacement cross section σ and the damage function v is established by conditional probability $P(\sigma,T_m|T)$, instead of using an empirical analytical relation $\sigma = \int_0^{T_m} v(T)d\sigma(T)dT$ where $d\sigma(T)$ is the differential recoil energy cross section or the probability of producing a recoil of energy between T and $T+dT$. The range of energy in the experiments is from 20eV to 130eV.

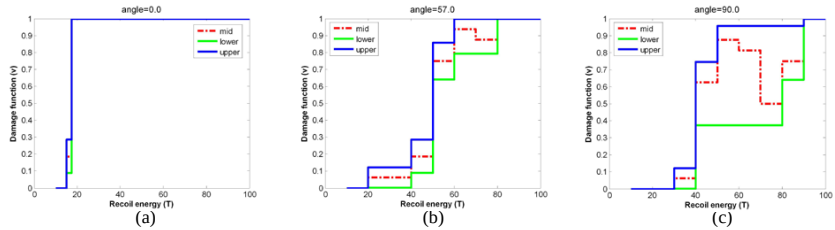

Figure 2. The example cumulative probability distributions of the generation of a stable Frenkel pair at different recoil energy (eV) and angles by MD simulation.

Table I. Examples of correspondence between the displacement cross section σ (barns) and the resistivity change rates $\Delta\rho/n$ (10^{-26} Ωcm/(e^-/cm^2)) (with minimum and maximum values) at each level of maximum transmitted energy T_m (eV) in a head-on collision [10]

<100>				<111>			
T_m	σ	$\Delta\rho/n$: *min*	$\Delta\rho/n$: *max*	T_m	σ	$\Delta\rho/n$: *min*	$\Delta\rho/n$: *max*
70	23.1578947	7.3366283	8.21892725	70	29.0789474	9.25116666	9.55835715
100	32.8947368	9.51916799	10.4014669	100	39.2105263	12.8622778	13.1694683

To validate the MD simulation model by the experiments that are conducted at macroscopic scale, we can compare the prior probability of model parameters θ in MD $P(\theta)$ and the posterior probability $P(\theta|\Delta\rho/n)$, calculated as

$$\mathbf{p}(\theta \mid \Delta\rho/n) = \frac{\mathbf{p}(\theta)\int\cdots\int\left[\mathbf{p}(\Delta\rho/n \mid \sigma,T_m)\mathbf{p}(\sigma,T_m \mid T)\mathbf{p}(T \mid \theta)\right]dTd\sigma dT_m}{\mathrm{dual}\int\cdots\int\left[\mathbf{p}(\Delta\rho/n \mid \sigma,T_m)\mathbf{p}(\sigma,T_m \mid T)\mathbf{p}(T \mid \theta)\mathbf{p}(\theta)\right]dTd\sigma dT_m d\theta} \quad (2)$$

As an illustration, a simple numerical example is shown as follows. Given a set of MD simulation model parameter θ, the interval probability that a Frenkel pair will be generated at an energy level less than 70 eV at direction <111> is estimated from Fig.2(b) as $\mathbf{P}(T\leq 70|\theta)=[0.7923,1]$. From LCC, we know $\mathbf{P}(T>70|\theta)=[0.2077,0]$. Additionally, $\mathbf{P}(T\leq 70|\sim\theta)=[0,1]$ and $\mathbf{P}(T>70|\sim\theta)=[1,0]$ since there is a lack of knowledge when the model parameters are different.

From Table I and the assumption of uniform distribution of $\Delta\rho/n$, we have $\mathbf{P}(\Delta\rho/n\leq 9.4048|\sigma=29.0789,T_m=70)=[0.5,0.5]$, $\mathbf{P}(\Delta\rho/n>9.4048|\sigma=29.0789,T_m=70)=[0.5,0.5]$, $\mathbf{P}(\Delta\rho/n\leq 9.4048|\sigma=39.2105,T_m=100)=[0,0]$, and $\mathbf{P}(\Delta\rho/n>9.4048|\sigma=39.2105,T_m=100)=[1,1]$. We do not have any information about what we did not measure. That is, with total ignorance, we have $\mathbf{P}(\Delta\rho/n\leq 9.4048|\sigma\neq 29.0789,T_m\neq 70)=[0,1]$, $\mathbf{P}(\Delta\rho/n>9.4048|\sigma\neq 29.0789,T_m\neq 70)=[1,0]$, $\mathbf{P}(\Delta\rho/n\leq 9.4048|\sigma\neq 39.2105,T_m\neq 100)=[0,1]$, and $\mathbf{P}(\Delta\rho/n>9.4048|\sigma\neq 39.2105,T_m\neq 100)=[1,0]$. Furthermore, $\mathbf{P}(\sigma=29.0789,T_m=70|T\leq 70)=[0.01,1]$, $\mathbf{P}(\sigma\neq 29.0789,T_m\neq 70|T\leq 70)=[0.99,0]$,

$\mathbf{P}(\sigma=39.2105, T_m=100|T\leq 70)=[0,0.01]$, $\mathbf{P}(\sigma\neq 39.2105, T_m\neq 100|T\leq 70)=[1,0.99]$,
$\mathbf{P}(\sigma=29.0789, T_m=70|T>70)=[0,0.01]$, $\mathbf{P}(\sigma\neq 29.0789, T_m\neq 70|T>70)=[1,0.99]$,
$\mathbf{P}(\sigma=39.2105, T_m=100|T>70)=[0.01,1]$, $\mathbf{P}(\sigma\neq 39.2105, T_m\neq 100|T>70)=[0.99,0]$.

Suppose the prior probability is $\mathbf{P}(\theta)=[0.4,0.5]$, the posterior probability after an instance of $\Delta\rho/n\leq 9.4048$ is observed is $\mathbf{P}(\theta|\Delta\rho/n\leq 9.4048)=[1,0.5]$ based on Eq. (2). The value of posterior probability increases compared to the prior. However, because of the lack of knowledge, the width of the interval probability also increases. Interval probability provides the extra information of how significant the epistemic component of uncertainty plays in the model validation. The slight overlap between the two intervals does not invalidate the model. Further experiments are needed, if more robust conclusions are desirable. As more knowledge is obtained, the posterior probability will converge to the classical precise one with the interval width gradually reduced.

Concluding Remarks

The GHMM and generalized interval Bayes' rule improve the robustness of the model validation process, where measurement data have inherent systematic errors and computational models contain intrinsic model errors. During validation, interval-valued posterior probability distributions of model parameters or model predictions are updated with the collected data from multiple scales or domains based on the generalized interval Bayes' rule. Validation is done by comparing the interval posterior probability with the prior. A substantial difference between the two invalidates the model. The epistemic uncertainty component during the inference provides us extra information so that more robust conclusions can be obtained.

Acknowledgements

This research is being performed using funding received from the DOE Office of Nuclear Energy's Nuclear Energy University Programs.

References

[1] Dempster A. (1967) Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, **38**(2):325-339

[2] Shafer G.A. (1990) Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ.

[3] Walley P. (1991) Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London.

[4] Ferson S., Kreinovich V. Ginzburg L., Myers D.S., and Sentz K. (2003) Constructing probability boxes and Dempster-shafer structures. *Sandia National Laboratories Technical report* SAND2002-4015, Albuquerque, NM.

[5] Wang Y. (2010) Imprecise probabilities based on generalized intervals for system reliability assessment. *International Journal of Reliability & Safety*, **4**(4): 319-342

[6] Kaucher E. (1980) Interval analysis in the extended interval space IR. *Computing Supplementa*, Vol.2, 33-49

[7] Wang Y. (2010) Multiscale uncertainty quantification based on a generalized hidden Markov model. *Journal of Mechanical Design*, **133**(3): 031004

[8] Babuška, I., Nobile, F., and Tempone, R. (2008). A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria. Computer Methods in Applied Mechanics and Engineering, **197**(29), 2517-2539.

[9] Oden, T., Moser, R., and Ghattas, O. (2010). Computer predictions with quantified uncertainty, Part I. SIAM News, **43**(9), 1-3.

[10] Vajda, P. (1977). Anisotropy of electron radiation damage in metal crystals. Reviews of Modern Physics, **49**(3), 481.