Chemical Physics Letters 499 (2010) 1–8

Contents lists available at ScienceDirect

Chemical Physics Letters

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

FRONTIERS ARTICLE

Uncertainty quantification: Making predictions of complex reaction systems reliable

Trent Russi, Andy Packard *, Michael Frenklach **

Department of Mechanical Engineering, University of California, Berkeley, CA 94720-1740, USA

ARTICLE INFO

Article history:
Received 1 September 2010
In final form 2 September 2010
Available online 9 September 2010

ABSTRACT

There is increasing need to make chemical reaction models and modeling more predictive. We examine the modeling methodology from the perspective of propagation of uncertainties, those in assumed model parameters along with those in experimental observations. Accepting the length of the uncertainty interval in the predicted property as a measure of model predictiveness, we examine methodological factors affecting it. Employing the recently introduced technique of Data Collaboration, we show that even 'harmless' assumptions, invoked explicitly or implicitly to alleviate a burden of numerical procedures, could lead to substantial differences in model predictiveness. We also demonstrate that the direct, one-step methodology, such as Data Collaboration, necessarily makes modeling more predictive and thus more reliable than a two-step approach typical of most current methods.

1. Introduction

1.1. Predictive models

The present societal concerns – global climate change, natural disasters, diverse energy sources, conservation, security, and so on – advocate the science and technology required to develop the ability of making reliable predictions. Undoubtedly, the road to attain this goal lies through building predictive models. Considering the complex nature of the phenomena in question, the working hypothesis for the scientific method is to gain fundamental understanding of the underlying processes. The knowledge gained through increasingly sophisticated experimentation and theory then forms the basis for the development of models. But what makes such models truly predictive?

The theoretical dream is to build models entirely from first principles. However, even most fundamental of the present models include uncertainties. There are many sources of model uncertainties: incomplete knowledge of the physical phenomena, truncated expansions of numerical methods, numerical diffusion, and the like. One view of model predictiveness is to gain higher and higher veracity for all parts of the model and by this virtue alone acquire trust in predicted results. Even in this possibly utopian view, the question of how to judge that the model predictions are sufficiently accurate needs to be answered. Furthermore, one would like to have a direction for advancing the model predictiveness. Not all model parts and not all of their uncertainties contribute equally to the accuracy of model predictions, especially when one is interested in a specific set of conditions. The usually complex, nonlinear nature of models of physical phenomena prevents one from identifying the extent to which individual uncertainties influence model predictions without analysis.

A broad field of study and techniques related to the numerical characterization of uncertainty has been termed *uncertainty quantification* (UQ). Its main objective is the propagation of a model's uncertainties to the model's prediction, and hence 'uncertainty quantification' has become synonymous with 'predictive modeling'. We here define 'predictive' to mean that the numerical result of a model is accompanied by its rigorously determined uncertainty bounds and a more *predictive* model is the one with more narrowly bounded interval.

1.2. Numerical approaches to uncertainty quantification

There are two fundamental frameworks to quantification of model uncertainty: probabilistic and deterministic. The former, which is currently more popular, begins with assuming prior distributions for parameter and/or experiment uncertainties, builds the analysis on Bayes' theorem and maximum likelihood principle, employs stochastic sampling of the system's deterministic model (for kinetics, for instance, repeatedly solving a system of ordinary differential equations), and as a result arrives at the posterior distributions of model parameters and model prediction(s) [1,2]. While providing a sound mathematical underpinning, a rigorous Bayesian framework for validation of computer models [3] is computationally expensive. Recent approaches designed to speed up the calculations without sacrificing much of the rigor demonstrate application to a single, overall reaction [4] or a small reaction set [5].

The same problem can be posed and solved using a deterministic framework, in which one characterizes uncertainties with hard

* Corresponding author.
** Corresponding author.
E-mail addresses: apackard@berkeley.edu (A. Packard), frenklach@berkeley.edu (M. Frenklach).

0009-2614/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2010.09.009

bounds rather than probability distributions. One such methodology, Data Collaboration, was introduced recently (see Section 2): given the documented bounds on experimental uncertainty and prior bounds on model parameters, one solves for bounds on model prediction(s). The key advantages of Data Collaboration are the computational efficiency and favorable scalability with the size of the problem. It is accomplished by a priori identification of smaller sets of active variables for individual responses, building quadratic response surfaces in those active variables, setting the problem as a set of quadratic inequalities, and solving the set with efficient numerical algorithms. The details are given in the next section. Here it is pertinent to mention that at least in one test study [6], the predicted intervals of Data Collaboration were consistent with posterior quantiles of the full Bayesian analysis with uniform priors.

1.3. Prediction intervals

While the differences and similarities between the Bayesian and Data Collaboration frameworks are perhaps interesting subjects for future studies, here we explore the opportunity to gauge the influence of approximations in uncertainty quantification methods on the accuracy of these methods predictions. Data Collaboration enables economical analysis of large-size systems, and we have demonstrated the technique using a 102-variable, 77-experiment example. One of the central features of Data Collaboration is the prediction on the feasible set, the set of parameter values constrained by the bounds of model parameters and by the bounds of training-set experiments. It is very hard to visualize the 102-dimensional region that constitutes the feasible set, yet it is this region that ultimately determines the length of the interval of the model prediction.

Various indicators point to an irregular geometric shape of the 102-dimensional region holding the feasible set [7,8]. The present study investigates the numerical effect of approximating the shape of this feasible region on the length of the prediction interval. Approximations involved in various UQ methods are rendered into different geometric shapes. For instance, linearizing model responses leads to ellipsoidal contours of posterior distribution. If we presume a geometric analogy between posterior quantiles of the probabilistic framework and the feasible set of the deterministic approach, our results assess accuracy of different UQ methods.

1.4. Reaction models

While neither our approach nor discussion are limited to it, our demonstration platform here is the kinetics of complex chemical reaction systems, those now common in the fields of combustion, atmospheric phenomena, astrophysics, material synthesis, and system biology. Physical models in these fields are complex networks of chemical species related to each other by molecular transformations. Their mathematical representation takes the form of a system of ordinary differential equations. The equations are highly nonlinear and the system, generally, does not have a closed-form solution. We assume here that the underlying chemical reaction network and hence the differential equations composed of reaction rate terms are known (or 'tentatively entertained' [9]). The model uncertainty is thus entirely associated with the insufficient knowledge of rate-law parameters. There are also uncertain inputs, yet we assume, to simplify the presentation, that the assumption of 'uncertain parameters' covers them all.

The uncertainty in model parameters usually originates from fitting calibration experiments that themselves have measurement uncertainties. In recent practice, rate coefficients are often being determined also by theory, quantum-chemical calculations of potential energies combined with rate-theory calculations. In the context of the present discussion, such can be qualified also as 'experiments' (computer experiments in this case) whose 'measurement' uncertainty comes from the approximation of the Hamiltonian in quantum-chemical calculations and theoretical assumptions employed by rate theories.

1.5. Paper outline

In this article, we approximate the feasible set in several ways and examine how these approximations affect uncertainty in model prediction. Section 2 describes our methodology of Data Collaboration. In Section 3 we describe the GRI-Mech 3.0 dataset and its feasible set, the primary example for our analysis. Several feasible set approximations are built in Section 4 and exercised in Section 5. We conclude with implications of our results to model and modeling being predictive.

2. Data collaboration

Data Collaboration is a mathematical framework for using theoretical models and experimental data from different but related sources to explore their collective information content. The methodology tests consistency among data and models [8], explores sources of inconsistency [8], discriminates among differing models [10], makes model interval predictions [7,11], and analyzes sensitivity of uncertainty propagation [12]. Applications to date of the Data Collaboration methodology include combustion science [7,8,12,13] and engineering [14], atmospheric chemistry [15], and system biology [10,16,17].

The Data Collaboration framework begins with the definition of the collective data, which we call a *dataset*.

2.1. Dataset

The basis of Data Collaboration is composed of an underlying physical process and associated model, a collection of experimental observations with respective uncertainties, algebraic surrogate models (response surfaces) representing parametric dependence of the physical-model predictions of the experimental observables on the uncertain parameters, and specialized constrained optimization algorithms.

In more detail, the components are:

- *Underlying physical process and corresponding model*: A complex physical process for which the research community wishes to build a predictive modeling capability (e.g., a chemical kinetics model for the combustion of natural gas). Appropriate conservation laws are known, but uncertain parameters, whose values are denoted by an n-dimensional parameter vector $\mathbf{x}$, limit predictability. The true value x_i of each individual parameter is bounded by expert-assessed uncertainties of the form $x_{i,\min} \leqslant x_i \leqslant x_{i,\max}$, defining a *prior-knowledge* hypercube $\mathscr{H}$ (e.g., a tabulation of rate coefficients assessed by a panel of experts). The model is a mathematical formulation of the physical process depending on these parameters (e.g., the parametrized ODEs of the chemical kinetics model), notated as $\phi(\mathbf{x})$.
- *Experimental observations and respective uncertainties*: Individual researchers propose, construct and carry out diverse, high-quality experiments involving the complex physical process. The reported measured outcome of observable Y is $d_{\text{lower}} \leqslant d \leqslant d_{\text{upper}}$. The observables identified for analysis are referred to as *targets*.
- *Physically-based models of experimental observations*: Individual experiments involve the complex physical process, along with the specific physical manifestation of the experimental investi-

gation (geometric, thermal, etc.). In other words, an individual experiment exercises the same underlying process (e.g., reactions oxidizing methane into carbon dioxide and water) but in differing physical environments (shock tubes, laminar flame burners, etc.). A physically-based model that predicts the measured observable, Y, is the mathematical composition of the process model, $\phi(\mathbf{x})$, and the specific (to Y) experiment model. At different parameter values $\mathbf{x}$, the physically-based model predicts, in general, different values of the measured observable.

- *Surrogate models of experimental targets*: Using simulation and computer experimentation, along with curve-fitting techniques, the outputs of the physically-based models are represented as algebraic functions of *active* parameters, those with significant effects. To date, we have specifically used quadratic polynomial and rational quadratic (quotients of quadratics) as our surrogate function classes. Mathematically, the surrogate model M closely approximates the physically-based model over $\mathscr{H}$.

In summary, an observable, Y, is both experimentally measured and predicted by a model. $M(\mathbf{x})$ is a model predicting Y, while d is the measured value. The discrepancy between the measurement and its model prediction is bounded by l from below and by u from above. The inequality $l \leqslant M(\mathbf{x}) - d \leqslant u$ combines the experimental and modeling information into a single set of constraints. The triple of {measurement d, uncertainty l and u, model M} is referred to as a *dataset unit*. A collection of dataset units, indexed by $e = 1,2,\ldots,m$, is a *dataset*, D. In other words, a dataset is a collection of all experimental targets selected for analysis, with all their respective uncertainties and surrogate models.

2.2. Feasible set

The subset of the parameter prior-knowledge hypercube $\mathscr{H}$ satisfying $l_e \leqslant M_e(\mathbf{x}) - d_e \leqslant u_e$ for all dataset units is the *feasible set*, $\mathscr{F}$. It expresses the collective constraints imposed presumably by all experimental and theoretical knowledge on the system.

Two natural UQ questions arise:

- *Consistency*: Is the dataset consistent, namely is $\mathscr{F}$ nonempty? In other words, is there at least one combination of parameter values, each within its respective uncertainty bounds, for which all of the model predictions are within their respective bounds?
- *Prediction*: Assuming consistency, what is the prediction of an unmeasured (but modeled) property, Y_P, in the form of a prediction interval that bounds the set of values M_P can take over $\mathscr{F}$? In other words, what is the interval predicted by the model for the yet undetermined property that satisfies all the available parameter and experiment data?

In developing computational schemes to answer these, questions of representation and approximation become relevant:

- *Representation*: Assuming consistency, how complex is $\mathscr{F}$, and are there economical/efficient manners to describe/approximate it, beyond the list of constraints that define it? Are these approximations more amenable to analysis than the defining description?

In this paper, we focus on how approximations in representation impact the predictive power of the dataset. We show that for high dimension problems, common in kinetics modeling, even if the model functions M_e are quadratic, a great deal of predictive power can be lost when approximations to $\mathscr{F}$ are used.

In the next Section (2.3) we discuss constrained optimization, noting that prediction involves both a maximization and minimization of M_P over $\mathscr{F}$. Other important questions (but not a focus of this paper) can be posed in the context of constrained optimization as well, including

- *Relevance*: How does the addition of a new dataset unit further constrain $\mathscr{F}$, and is this new constraint useful toward answering the scientific questions of the research community?
- *Sensitivity*: How sensitive are these various answers to each individual dataset unit's values and uncertainties?

These are addressed in detail in Refs. [10,12,16,18].

2.3. Optimization on the feasible set

The prediction problem (for example) of Data Collaboration involves minimization and maximization of a model over the feasible set $\mathscr{F}$. The set $\mathscr{F}$ is defined by a collection of inequality constraints, namely those defining the parameter hypercube $\mathscr{H}$ along with the many individual constraints $l_e \leqslant M_e(\mathbf{x}) - d_e \leqslant u_e$ of each dataset unit. In that context, we briefly review some key results pertaining to constrained optimization.

Without loss of generality, we discuss *minimizations* with inequality constraints, referred to as primal problems,

$$\begin{aligned} &p^* = \min_{x\in\mathbb{R}^n} f(x),\\ &\text{subject to}: \ g_j(x) \leqslant 0,\\ &\text{for } j = 1,\ldots,m. \end{aligned}$$

Direct attempts at this constrained optimization typically only yield an upper bound to p^*, since the true minimum may not be found, due to nonconvexity of the function and/or constraint set, for example. In order to bracket p^*, a lower bound is also needed. Toward that goal, the associated dual problem

$$\begin{aligned} &q^* = \max_{\lambda\in\mathbb{R}^m} \min_{x\in\mathbb{R}^n} f(x) + \lambda^{\mathrm{T}} g(x),\\ &\text{subject to}: \ \lambda_j \geqslant 0,\\ &\text{for } j = 1,\ldots,m, \end{aligned}$$

always has $q^* \leqslant p^*$, giving (if q^* is reliably computed) a lower bound to the primal. As we are considering minimization, lower and upper bounds on p^* are referred to as outer and inner bounds, respectively. Additionally, the solution of the dual problem informs how p^* is affected by changes in the constraints. Specifically, the primal problem with variable constraints, v,

$$\begin{aligned} &p^*(v) = \min_{x\in\mathbb{R}^n} f(x),\\ &\text{subject to}: \ g_j(x) \leqslant v_j,\\ &\text{for } j = 1,\ldots,m, \end{aligned}$$

is related to the optimal dual variables λ^*, which act as global (one-sided) sensitivities through the bound

$$q^* - v^{\mathrm{T}}\lambda^* \leqslant p^*(v) \quad \text{for all } v \in \mathbb{R}^m.$$

Finally, and of critical importance, if the functions f and all g_j are quadratic (not necessarily sign-definite), then the dual problem is solved very efficiently via semi-definite programming (SDP) [19,20]. Problems with hundreds of variables (n) and constraints (m) are routine, even in a desktop environment. Taken together, the efficiently computed bound q^* and sensitivities λ^* highlight an important consequence for uncertainty quantification with quadratic response surface models.

Although our deterministic perspective on uncertainty quantification directly gives the constrained minimization (primal) discussed here, we note that under the quadratic-dependence restriction on f and all g_j, the dual problem is actually equivalent to a stochastic formulation [21]. Namely, replace the deterministic

variable x with a random variable X (restricted only to have finite variance) and modify the cost and constraints to reflect mean values, as

$$s^*(v) = \min_X \mathbf{E}[f(X)],$$
$$\text{subject to : } \mathbf{E}[g_j(X)] \leqslant v_j,$$
$$\text{for } j = 1, \ldots, m,$$

where $\mathbf{E}$ denotes expectation and the minimization is taken over all random variables X with finite variance. Then $s^*(v) = q^*(v)$, where $q^*(v)$ refers to the dual problem for $p^*(v)$, and hence p^* and s^* satisfy global sensitivity relations $s^*(0) - v^T\lambda^* \leqslant s^*(v) \leqslant p^*(v)$ for all $v \in \mathbb{R}^m$. So, while s^* is not a Bayesian estimate as discussed in Section 1.2, it nevertheless has appeal for those inclined towards probabilistic analysis. Furthermore, s^* (along with global sensitivities) is computed efficiently, it is independent of priors, and satisfies a known relation to the deterministic minimization.

2.4. Dataset consistency

As discussed above, the feasible set is a representation of the complete collaborative information contained in a dataset and questions in the Data Collaboration framework are posed as optimization problems over the feasible set. The first question is that of the dataset consistency. To assess it numerically, we introduce the *consistency measure* which answers the question 'What is the largest percentage of uncertainty reduction such that there exists a feasible parameter vector?' Associated with a given dataset D, it is notated C_D and posed as an optimization problem,

$$C_D := \max_{\gamma,\mathbf{x}} \gamma, \quad \text{subject to :}$$
$$\mathbf{x} \in H,$$
$$(1-\gamma)l_e \leqslant M_e(\mathbf{x}) - d_e,$$
$$M_e(\mathbf{x}) - d_e \leqslant (1-\gamma)u_e,$$
$$\text{for } e = 1, \ldots, m.$$

In this definition, the original constraints $l_e \leqslant M_e(\mathbf{x}) - d_e \leqslant u_e$ are augmented with a scalar γ, where positive values of γ imply tightening of the constraint, and negative values imply loosening. The consistency measure quantifies how much the constraints can be tightened while still ensuring the existence of a set of parameter values whose associated model predictions match (within bounds) the experimental targets. The dataset is *consistent* if the consistency measure is nonnegative, and is *inconsistent* otherwise. Inconsistency simply means that at the accepted uncertainty levels, there is no parameter set $\mathbf{x}$ that satisfies all of the parameter and experiment constraints.

2.5. Model prediction

Given a set of conditions not exercised experimentally but with the property P predicted by model M_P, what is the range of values this model computes over the domain of feasible parameter values? The DC computation of this range is posed as two optimization problems for the lower and upper interval endpoints, L_P and U_P,

$$L_P := \min_{\mathbf{x}\in\mathscr{F}} M_P(\mathbf{x}), \tag{1}$$
$$U_P := \max_{\mathbf{x}\in\mathscr{F}} M_P(\mathbf{x}). \tag{2}$$

The length $U_P - L_P$ quantifies the amount of uncertainty in M_P's value conditioned that the M_P's parameters are in the feasible set $\mathscr{F}$. Figure 1 illustrates in a graphical form the interval (L_P, U_P) predicted on a feasible set, $\mathscr{F}$, and its comparison to the interval predicted on the entire prior-knowledge domain $\mathscr{H}$.

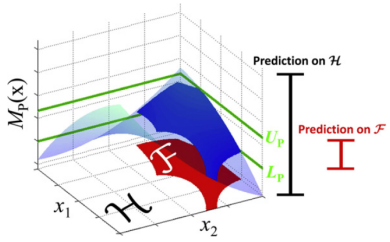

Figure 1. Illustration of the prediction on a feasible set.

2.6. Connection to optimization

As outlined in Section 2.3, bounds to individually bracket C_D, L_P and U_P are solved efficiently using polynomial optimization techniques when all the models are quadratic in the parameters [7,11,20]. Further use of branch-and-bound techniques [10] tightens the bracketing interval of each quantity. Solving these optimizations also automatically generates sensitivities of the computed results to the parameter and experiment uncertainties [12]. Extension of the Data Collaboration methodology to nonquadratic models is treated in Ref. [16].

3. Example Dataset: GRI-Mech 3.0

The GRI-Mech 3.0 dataset is a collection of experimental observations (targets) and corresponding surrogate models selected and developed to study the chemical kinetics of pollutant formation in the combustion of natural gas [22]. This dataset consists of 77 targets and paired models. Each surrogate model is a (nonconvex) quadratic surrogate, expressing $\log_{10}$ of the target in terms of $\log_{10}$ of parameters (mostly pre-exponential factors of the Arrhenius reaction rate expressions, but also species enthalpies of formation), normalized to take values between -1 and $+1$. Altogether, the surrogate models involve 102 active parameters. The uncertainty associated with experimental measurements have been assigned by domain experts [22,23]. Further details can be found in Refs. [7,22,23].

Testing consistency, as described in Section 2.4, led to the discovery of an inconsistency in the GRI-Mech 3.0 dataset and identified as its main cause one of the targets, target F5, a laminar flame speed of a stoichiometric methane-air mixture at 4.9 atm [22], with the assessed range of experimental uncertainty of $[37.7, 41.7]$ cm/s [23]. Therefore, in the examples using the GRI-Mech 3.0 dataset, we removed the constraints associated with target F5. Later, in Section 5.2, target F5 is used in a prediction example.

Thus, for the purpose of the present work, the GRI-Mech 3.0 dataset is a collection of 76 dataset units, along with a 102-dimensional prior-knowledge hypercube $\mathscr{H}$ of parameters. In other words, a point in $\mathscr{H}$ is a set of 102 parameters (a 102-dimensional vector) each having a value within its prescribed interval of uncertainty.

The GRI-Mech 3.0 feasible set, $\mathscr{F}$, is a subset of $\mathscr{H}$; $\mathscr{F}$ contains only those combinations of the prescribed parameter values that predict all dataset targets each within its respective uncertainty bounds. Mathematically, feasible set $\mathscr{F}$ occupies a 102-dimensional volume, embedded within and constrained by the $[-1, 1]^{102}$ hypercube $\mathscr{H}$ and by nonconvex quadratic constraints

of the dataset units (several cross-sections are examined and displayed in Ref. [7]).

Due to this complexity, high-dimensionality and a large number of non-planar encasing surfaces, it is difficult to visualize or discern the geometry such as that of the GRI-Mech 3.0 feasible set. Furthermore, attempts to approximate it run into the well-known curse of dimensionality; e.g., estimating the volume of $\mathscr{F}$ through logical and seemingly close approximations exceeds by many orders of magnitude its actual volume. This brought us to question: What affect would innocently looking approximations to $\mathscr{F}$, such as those described in Section 4, have on the range of the model prediction? Answering this question constitutes the main subject of the present study.

4. Feasible set approximations

The feasible set as defined by the Data Collaboration methodology is in general nonconvex, high-dimensional, and bounded by many non-planar surfaces. It is natural to aspire to develop a good approximation to such a feasible set that is both easy to analyse and easy to describe. If found, a mathematically simple approximation could help with quantification of uncertainty. In fact, many of the uncertainty quantification methods explicitly or implicitly resort to such approximations.

To explore possible consequences of approximating the feasible set, we investigate several systematically constructed approximations. To help the reader in visualizing the general character of these approximations, we present in Figure 2 simple toy examples illustrating them.

The rest of this section describes methods used for generating the approximations. Afterwards, in Section 5, we explore how good these approximations are for the GRI-Mech 3.0 feasible set.

4.1. Sampling and principal components

Principal component analysis (PCA) is often used to account for the variability in a data set. By sampling a set and using PCA, we can get a sense of the basic ellipsoidal-like shape of the set. The first (largest by magnitude) principal component approximates the direction in which the set is the widest, and the last principal component approximates its thinnest direction (cf. Figure 2, PCA).

To sample a set defined by simple algebraic inequalities, we employed a randomized gas kinetics point generator (GKPG) [24] and a random walk [25] algorithms, based on a simple gas-dynamics simulation of a particle moving within the set. Collisions with other particles are modeled as instantaneous changes in direction (new direction is a unit-vector, selected from a uniform distribution on the unit-sphere) and reflections off walls (the active constraint) are lossless, satisfying Fresnel's law. As the simulation proceeds, the particle's location is periodically recorded, generating a sample of points from the set. A principal component analysis of these points, as shown in Figure 2, (PCA) is then used to generate two approximations described in Sections 4.2 and 4.3 below.

4.2. Rotated truncated hyperrectangle

The principal component directions found using the method in the previous section provide a natural rotated coordinate system for the set approximation. The smallest hyperrectangle, aligned in these rotated coordinates, and containing the set $\mathscr{F}$ is a convex, outer approximation of $\mathscr{F}$, described by $2n$ linear constraints. In fact, the constraints can be found using the Data Collaboration method for an interval prediction.

Let $\mathbf{p}$ be one of the principal component directions, an $n \times 1$ vector. The maximum and minimum values that $\mathbf{p}^{\mathrm{T}}\mathbf{x}$ can take over values of $\mathbf{x}$ in the feasible set lead to two of the desired constraints in the rotated hyperrectangle. This calculation is an interval prediction of the model $\mathbf{p}^{\mathrm{T}}\mathbf{x}$ with the Data Collaboration method. By repeating this procedure for each principal component direction, the rotated hyperrectangle is found.

Furthermore, the approximation of the feasible set can be improved by including the original parameter bounds (the bounds of $\mathscr{K}$). As a result, the approximation will be a truncated

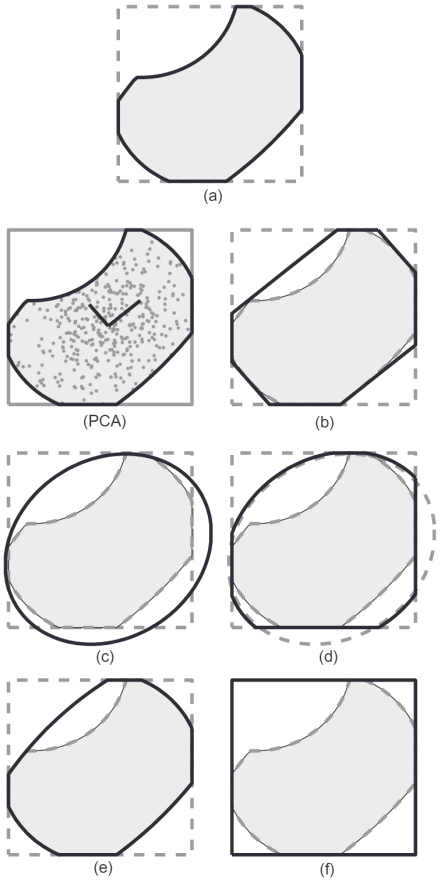

Figure 2. Approximations to a toy feasible set in two dimensions. In each case the relevant set is shown with a bold outline, the actual feasible set $\mathscr{F}$ is shaded, and the prior-knowledge region, $\mathscr{K}$, is shown as a square. Approximation formulations are described in the text. (a) The actual feasible set; (PCA) Principal component analysis of a feasible set sample resulting in the two directions shown, (b) the feasible set enclosed by $\mathscr{K}$ and lines parallel to the PCA directions, (c) the feasible set enclosed by the smallest ellipse, determined using the PCA directions and weights along with the mean of the sampling, without $\mathscr{K}$, (d) same as (c), but constraining to $\mathscr{K}$, (e) a convex outer approximation of the feasible set, as described in the text, and (f) $\mathscr{K}$ by itself.

hyperrectangle, more generally a polytope, with $4n$ linear constraints (cf. Figure 2b).

4.3. Bounding ellipsoid

Rather than a hyperrectangle (defined by n pairs of mutually orthogonal linear constraints), the set $\mathscr{F}$ can also be bounded using an ellipsoid, which is defined by a single quadratic constraint. Here, we choose both the principal components and their associated weights to define rotation and proportions of the bounding ellipsoid. The arithmetic mean, $\overline{\mathbf{x}}$, of the sample points is used to center the ellipsoid.

The ellipsoidal approximation to the feasible set is defined as

$$\{\mathbf{x} \in \mathbb{R}^n | (\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}} Q (\mathbf{x}-\overline{\mathbf{x}}) \leqslant \alpha\}, \tag{3}$$

where α is a scaling factor to be determined. We would like the smallest α such that the ellipsoid contains the feasible set. This can be calculated using the maximization half of the Data Collaboration prediction method,

$$\begin{aligned} &\alpha := \max_{\mathbf{x}} (\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}} Q (\mathbf{x}-\overline{\mathbf{x}}), \\ &\text{subject to} : \mathbf{x} \in \mathscr{F}. \end{aligned} \tag{4}$$

This ellipsoidal approximation is defined by a single quadratic constraint (cf. Figure 2c). As with the rotated hyperrectangle, the approximation can be improved by including the $2n$ parameter bound (prior information bound) constraints. This results in a truncated ellipsoid (cf. Figure 2d).

4.4. Convex outer approximation of a nonconvex quadratically constrained set

If the feasible set is defined by nonconvex quadratic constraints, then we can approximate it using convex quadratic constraints.

Let the feasible set be defined by p two-sided constraints, written as:

$$T_k \leqslant \mathbf{x}^{\mathrm{T}} \mathbf{Q}_k \mathbf{x} + \mathbf{b}_k^{\mathrm{T}} \mathbf{x} + c_k \leqslant R_k, \tag{5}$$

where $k = 1, \ldots, p$, T_k and R_k are scalars, and $\mathbf{Q}_k$, $\mathbf{b}_k$, and $\mathbf{c}$ are the quadratic, linear, and constant coefficients, respectively. The matrices $\mathbf{Q}_k$ can be decomposed as

$$\mathbf{Q}_k = \mathbf{Q}_{k,\mathrm{pos}} + \mathbf{Q}_{k,\mathrm{neg}}, \tag{6}$$

where $\mathbf{Q}_{k,\mathrm{pos}}$ is positive semi-definite and $\mathbf{Q}_{k,\mathrm{neg}}$ is negative semi-definite. For each value of k we solve the following two optimization problems using the Data Collaboration prediction methods,

$$\begin{aligned} &\mu_k := \max_{\mathbf{x}} \mathbf{x}^{\mathrm{T}} \mathbf{Q}_{k,\mathrm{pos}} \mathbf{x} + \mathbf{b}_k^{\mathrm{T}} \mathbf{x} + c_k, \\ &\text{subject to} : \mathbf{x} \in \mathscr{F}, \end{aligned} \tag{7}$$

$$\begin{aligned} &\omega_k := \min_{\mathbf{x}} \mathbf{x}^{\mathrm{T}} \mathbf{Q}_{k,\mathrm{neg}} \mathbf{x} + \mathbf{b}_k^{\mathrm{T}} \mathbf{x} + c_k, \\ &\text{subject to} : \mathbf{x} \in \mathscr{F}. \end{aligned} \tag{8}$$

The approximation simply replaces the nonconvex quadratic constraints Eq. 5 with convex ones,

$$\begin{aligned} &\{\mathbf{x} \in \mathscr{H} : \ldots \\ &\mathbf{x}^{\mathrm{T}} \mathbf{Q}_{k,\mathrm{pos}} \mathbf{x} + \mathbf{b}_k^{\mathrm{T}} \mathbf{x} + c_k \leqslant \mu_k, \ldots, \\ &\mathbf{x}^{\mathrm{T}} \mathbf{Q}_{k,\mathrm{neg}} \mathbf{x} + \mathbf{b}_k^{\mathrm{T}} \mathbf{x} + c_k \geqslant \omega_k, \ldots, \\ &\text{for } k = 1, \ldots, p\}. \end{aligned} \tag{9}$$

and is shown in Figure 2e. This method of approximating the feasible set, like all the rest described in this section, can produce poor approximations in high dimensions. It is used here only for the purpose of comparison to the rigorous methods of Data Collaboration.

5. Results and discussion

The approximations developed in the previous section are now compared to the original feasible set. We perform two tests, both carried out with the GRI-Mech 3.0 dataset described in Section 3. The first test is direct volume sampling aimed at evaluating the volume ratios of the feasible set and its approximations. In the second test, we use the methodology of Data Collaboration to compare the lengths of the prediction interval computed for target F5.

5.1. Volume sampling

Consider a set, A, to be an approximating set that contains the true feasible set, $\mathscr{F}$. Sample points in A and reject those not in $\mathscr{F}$ (i.e., a rejection sampling of the feasible set). The fraction of not rejected to the total number of trials estimates the ratio of the volume of $\mathscr{F}$ to the volume of A. If this volume ratio is close to 1, we expect set A to be a good approximation of the feasible set $\mathscr{F}$.

In each of the five cases, 7.5×10^8 points were sampled from the approximating set. The GKPG algorithm described in Section 4.1 was used to sample the approximations developed in Section 4: truncated rotated hypercube (cf. Figure 2b), ellipse (cf. Figure 2c), truncated ellipse (cf. Figure 2d), and convex quadratic approximation (cf. Figure 2d), all encasing the actual feasible set, $\mathscr{F}$ (cf. Figure 2a). In the last case, we sampled uniformly the volume of $\mathscr{H}$ itself (cf. Figure 2f).

In each of these cases, not a single point was found to be in the feasible set! Numerically, the volume ratio is lower than $1/7.5 \times 10^8 = 1.3 \times 10^{-9}$, implying that the approximating sets are rather poor representations of the actual feasible set, as far as volume is concerned.

We recall that the GRI-Mech 3.0 feasible set and its approximations are in 102 dimensions. Consequently, the volume sampling results are a clear manifestation of the curse of dimensionality. With the increase in dimensionality, most of the volume (mass) resides at the space boundary. For instance, decreasing the radius of a 102-dimensional sphere by 1% lowers its volume by about 64%. As another relevant comparison, the volume ratio of a 102-dimensional sphere to its encasing hypercube is about 3×10^{-72}.

The results of the volume sampling illustrate that what appears in two or three dimensions as small approximations to feasible set boundaries turn into huge differences in high dimensions, and it is high-dimensional space of parameters which is characteristic of complex reaction models. The significance of the higher volume of the set approximating the feasible set is that many more parameter combinations become allowable. The next test illustrates the consequence of the increase in the allowable space on evaluated uncertainty in model prediction.

5.2. Interval prediction on the feasible set

One of the most significant features of the Data Collaboration methodology is rigorous evaluation of the model prediction *interval* by computing its maximum and minimum values over the feasible set of parameters. As described in Section 2, in Data Collaboration we are not necessarily concerned with identification of the feasible set itself, but with computing variation of various properties over all its points. The methodology allows us to accomplish this in most rigorous manner, without invoking approximations to the feasible set. We employed this methodology to compare the length of the prediction interval computed over the feasible set to that obtained with the feasible set replaced by its approximations.

We performed computations for the GRI-Mech 3.0 target F5, a laminar flame speed of a methane-air mixture [22]. As indicated

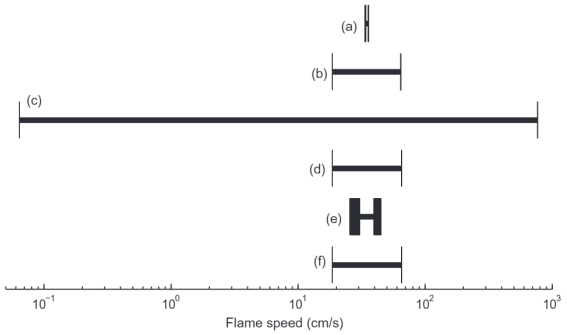

Figure 3. Interval prediction of GRI-Mech 3.0 target F5 using (a) the actual feasible set, $\mathscr{F}$, (b) a rotated and truncated hyperrectangle, (c) an ellipse, (d) a truncated ellipse, (e) convex quadratic approximation, and (f) the prior-knowledge hypercube, $\mathscr{K}$ (cf. Figure 2). The thickness of the bounding vertical lines indicates the differences in inner and outer bound predictions.

in Section 3, the 'documented' uncertainty of this target was found to be inconsistent with the rest of the targets, and therefore excluded from the dataset. Now, we compute the interval of predictions for target F5 based on the 76-target GRI-Mech 3.0 dataset.

Carrying out the Data Collaboration calculations with the actual feasible set, $\mathscr{F}$, resulted in a prediction interval with outer bounds of [33.65, 35.81] and inner bounds of [34.07, 35.61] cm/s. For comparison, we repeated the calculations replacing the feasible set with its approximations: the truncated rotated hyperrectangle, the bound ellipse, the truncated bounding ellipse, the convex quadratic approximations, and the prior-knowledge bounds. The obtained results are displayed in Figure 3. As expected, the prediction over the actual feasible set gives the tightest bounds. All other approximations result in significantly larger prediction intervals. Among them, the simplest approximation of the feasible set, the bounding ellipse, yields the largest prediction interval, over-predicting by about a factor of 400. This outcome is explained by the fact that the approximating ellipsoid overextends beyond $\mathscr{K}$ (cf. Figure 2c).

The dramatic differences seen in the prediction ranges are not the result of numerical artifacts or some peculiarity of the target F5's model. By contrast, suppose the experimental uncertainty in all 76 dataset units is increased by 10%, resulting in an expanded feasible set. On this expanded set, the predicted range for target F5 is bracketed by outer bounds [28.7, 40.2]. This is wider than the nominal prediction, as expected, but still far narrower than the ranges predicted with approximations tested above.

These results reiterate the message of the volume sampling on the poor quality of the approximations. More importantly, they reveal the extent that the seemingly 'harmless' approximations impact the key practical aspect – model prediction and hence reliability of modeling. Particular troublesome is the excessive over-prediction by the ellipsoidal approximation, as it is the most common approximation invoked explicitly or implicitly in numerical approaches to uncertainty quantification.

6. Implications to predictive modeling

A model with a smaller range of uncertainty computed for a desired property can be said to be more predictive or, in other words, more reliable. Obviously, many 'physical' factors affect this, like knowledge of the physical process in question, authenticity of the physical-model, validity of computational procedures, and accuracy of parameters. It is imperative, however, to be able to compute the uncertainty in the predicted property correctly, without introducing additional errors due to numerical methods.

The current paradigm of modeling, and certainly in the field of reaction kinetics, is a two-step process. In the first step, the model parameter values (rate coefficients, species thermodynamics, transport properties) are determined either through 'calibrating' experiments or quantum-chemical theory. The outcome is a compilation of these parameter values, ideally with uncertainty bounds specified. In the second step, a system of differential equations are solved for the conditions of interests, employing the best-available nominal parameter values. The sought-after objective is to propagate the documented parameter uncertainties through the differential equation integration.

This two-step process is represented in the present study by the test case with the prior-knowledge, $\mathscr{K}$. Indeed, performing a calculation on the basis of the parameter compilation presumes that all the values within the uncertainty bounds of each and every parameter are allowable. If the objective of the computation is to predict a property, then this constitutes a prediction on $\mathscr{K}$. In contrast, in Data Collaboration uncertainties of both parameters and experimental targets are transfered to model prediction directly, in a single step – this is the meaning of 'prediction on the feasible set, $\mathscr{F}$'. (A similar, single step uncertainty 'propagation' can also be accomplished with a full Bayesian approach [3]).

Unmistakably, $\mathscr{K}$ exceeds by volume the actual feasible set, $\mathscr{F}$ (cf. Figure 2f). Also, as shown in the test performed in Section 5.2, the prediction interval on $\mathscr{K}$ greatly exceeds the prediction interval on $\mathscr{F}$ (by a factor of about 20 in that specific test, Figure 3f). This indicates that the direct, one-step methodology necessarily results in more predictive modeling than a two-step approach.

Further increase in the predictive power of a model and modeling is to stay as close as possible to the true feasible set. The results of the tests performed here demonstrated that seemingly close approximations come with consequences that are amplified in high dimensions. Current reaction models of practical importance have on the order of $10^2 - 10^3$ reaction steps and hence their $\mathscr{K}$ space has a very large dimension. In such cases, even 'harmless' assumptions, invoked explicitly or implicitly to alleviate a burden of a numerical method, will lead to substantial differences in model predictiveness.

Acknowledgment

The work was supported by the NSF Chemistry Division, Cyber-enabled Chemistry, Grant No. CHE-0535542.

References

[1] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, CRC, Boca Raton, FL, 2004.
[2] M.C. Kennedy, A. O'Hagan, J. Roy. Stat. Soc. B 63 (2001) 425.
[3] M.J. Bayarri et al., Technometrics 49 (2) (2007) 138.
[4] H.N. Najm, B.J. Debusschere, Y.M. Marzouk, S. Widmer, O.P.L. Maître, Int. J. Numer. Meth. Eng 80 (2009) 789.
[5] H. Cheng, A. Sandu, Math. Comput. Simul. 79 (2009) 3278.
[6] R. Feeley, A. Packard, M. Frenklach, R. Paulo, J. Sacks, Unpublished, 2006.
[7] M. Frenklach, A. Packard, P. Seiler, R. Feeley, Int. J. Chem. Kinet. 36 (1) (2004) 57.
[8] R. Feeley, P. Seiler, A. Packard, M. Frenklach, J. Phys. Chem. A 108 (44) (2004) 9573.
[9] G.E.P. Box, W.G. Hunter, Technometrics 7 (1) (1965) 23.
[10] R. Feeley, M. Frenklach, M. Onsum, T. Russi, A. Arkin, A. Packard, J. Phys. Chem. A 110 (21) (2006) 6803.
[11] M. Frenklach, A. Packard, P. Seiler, Prediction uncertainty from models and data, in: Proc. American Control Conference, IEEE, New York, Anchorage, Alaska, 2002, pp. 4135–4140.
[12] T. Russi, A. Packard, R. Feeley, M. Frenklach, J. Phys. Chem. A 112 (12) (2008) 2579.
[13] M. Frenklach, Proc. Combust. Inst. 31 (2007) 125.
[14] J.P. Spinti, B. Hochstrasser, P.J. Smith, Oxy-gas combustion for efficient CO_2 capture: Effect of near burner mixing on velocity and composition fields, North American Mixing Forum, Victoria, BC, Canada, 2010.
[15] G.P. Smith, M.Frenklach, R.Feeley, A.Packard, P.Seiler, A system analysis approach for atmospheric observations and models: the mesospheric ho_x dilemma, J. Geophys. Res. (Atmospheres) 111(D23301).
[16] R.P. Feeley, Fighting the curse of dimensionality: A method for model validation and uncertainty propagation for complex simulation models, Ph.D. thesis, University of California, Berkeley, CA, 2008.
[17] T.-M. Yi et al., Application of robust model validation using SOSTOOLS to the study of G-protein signaling in yeast, in: Proceedings of Foundations of System Biology in Engineering, 2005, pp. 133–136.
[18] M. Frenklach, A. Packard, R. Feeley, Optimization of reaction models with solution mapping, in: R.W. Carr (Ed.), Comprehensive Chemical Kinetics, Vol. 42, Elsevier, 2007, p. 243. Ch. 6.
[19] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004.
[20] P. Seiler, M. Frenklach, A. Packard, R. Feeley, Optim. Eng. 7 (4) (2006) 459.
[21] E. Frazzoli, Z. Mao, J. Oh, E. Feron, AIAA J. Guidance Control 24 (1) (2001) 79.
[22] G.P. Smith et al., GRI-Mech 3.0, http://www.me.berkeley.edu/gri_mech/.
[23] X. You, T. Russi, A. Packard, M. Frenklach, Proc. Combust. Inst. 33, in press.
[24] L.H. Lee, K. Poolla, Statistical validation for uncertainty models, in: Feedback Control, Nonlinear Systems, and Complexity, Vol. 202/1995, Springer, Berlin/Heidelberg, 1995, pp. 131–149.
[25] S. Vempala, Algorithmic convex geometry, http://www.cc.gatech.edu/vempala/acg/notes.pdf (2008).

Trent Russi received his B.A. in Film and Digital Media from the University of California at Santa Cruz in 2003, and his B.S. and Ph.D. in Mechanical Engineering from the University of California at Berkeley in 2004 and 2010, respectively. He is the recipient of the ASME Leonard Farber Award and UC Berkeley's Outstanding Graduate Student Instructor Award. Trent's doctoral work was under the advisement of Andy Packard and Michael Frenklach and was completed with the dissertation entitled 'Uncertainty Quantification with Experimental Data and Complex System Models'.

Andrew Packard joined the University of California at Berkeley (UCB) Mechanical Engineering Department in 1990. His technical interests include quantitative nonlinear systems analysis and optimization and data structure issues associated with large-scale collaborative research for predictive modeling of complex physical processes. He is an author of the Robust Control Toolbox distributed by Mathworks. The Meyersound X-10 loudspeaker utilizes feedback control circuitry developed by his UCB research group. He received the campus Distinguished Teaching Award, the 1995 Eckman Award, and the 2005 IEEE Control System Technology Award. He is an IEEE Fellow.

Michael Frenklach is Professor in the Department of Mechanical Engineering of the University of California at Berkeley. He received his Diploma in Chemical Technology from the Mendeleyev Chemical-Technological University, Moscow, Russia, and his Ph.D. in Physical Chemistry at Hebrew University, Jerusalem, Israel. His faculty appointments began in 1979 in the Department of Chemical Engineering at Louisiana State University. In 1985 he joined the Materials Science Department of the Pennsylvania State University, and in 1995 he accepted his current position at Berkeley. Professor Frenklach's research interests are primarily in the area of chemical kinetic modeling and elucidation of reaction mechanisms of complex reaction networks, including combustion chemistry, particulate matter, and carbonaceous materials, such as diamond and graphene. His current activities focus on global system approach to predictive modeling and uncertainty quantification.