The Heterogeneous Multi-scale Method: A Mathematical Framework for Multi-scale Modeling

Weinan E

Princeton University, Princeton, NJ 08544

Contents

1 Introduction

In recent years, multi-scale modeling has received a great deal of attention in many disciplines of science and engineering. While almost every problem that we encounter has some multi-scale features in one way or another, and multi-scale modeling has been used in various forms for a very long time, traditionally the predominant approach to scientific modeling is to focus on one particular scale, treating other scales either empirically, or in a very restrictive setting. Constitutive relations, for example, are created to model the effects of the smaller scales. They can be the "strike of a genuis", representing fundamental physical insight, or a convenient way of "sweeping things under the rug", hiding our basic ignorance about the process at the micro-scale. In other cases when microscopic modeling is used, one is often forced to study macroscopically very simple systems, such as homogeneous systems, since realistic macroscopic systems are beyond the reach of these methods.

The recent surge of interest on multi-scale modeling is due largely to the realization that effective modeling strategies can be found by considering different models at different scales simultaneously, leading to models or simulation methods that can handle realistic macroscopic systems and at the same time, faithful to the necessary microscopic details. Many strategies have been proposed for different multi-scale problems. Some are specific to the particular problem; but some are of a general nature. Among the general ideas that have been suggested are information or parameter passing, also referred to as sequential or serial coupling methods, domain decomposition, adaptive model refinement, and various coarse-graining techniques.

In light of this vast development involving many different scientific disciplines, it is appealing to set up a general mathematical framework that would ideally apply to different situations. The heterogeneous multi-scale method (HMM) was developed exactly for this purpose [7, 9]. Recent advances seem to suggest that HMM does hold the promise of providing general guidelines for designing multi-scale methods for many different problems in different areas, as well as a framework for analyzing the stability, accuracy and efficiency of multi-scale methods. The motivation is very similar to that of the finite element method: Even though most of the basic ideas of the finite element method had already existed in the 40's and 50's, later work by the mathematics community on the framework of the finite element method, including the weak formulation of the problems, setting up finite element spaces, and basic error estimates, were essential for the wide range of successes of the finite element method. While HMM certainly has not received the same level of acceptance as the finite element method, we do believe that it is indeed a very convenient framework, useful for many problems.

We will put off the discussion of HMM until the next section. But at this point, let us make two important remarks in order to put things into perspective. The first remark is that to a large extend, HMM is an abstraction and improvement of the ideas that existed in different forms in, for example, in the Car-Parrinello method [5],

the quasi-continuum method [22], kinetic schemes for gas dynamics [18], the work in [23] on stochastic ODEs with multiple time scales, the projective dynamics for stiff ODEs [12], etc. We have found both the abstraction and the improvement useful. The second remark is that while HMM is a useful framework, it does not solve any specific problem by itself, since it leaves out several important details that need to be filled out according to the particular problem that it is applied to. Again this is similar to the situation in finite element method: The finite element method is useless unless we know how to construct the finite element space for the particular problem. We will illustrate this through several examples.

In this paper, we wil focus on two important settings for which a multi-scale method is called for.

In the first case we are only interested in the macroscopic behavior of the system, but we do not have at our disposal a satisfactory macroscopic model that is convenient for numerical computations. Instead, we have a microscopic model that we trust.

In the second case, besides the macroscopic behavior, we also would like to know about some microscopic details, such as samples of typical microstructures, microscopic behavior near defects, and defect dynamics.

In both cases, we are NOT interested in the details of the microscopic states everywhere. Our main objective is to design simulation methods that are much more efficient than solving the microscopic model everywhere. Such a viewpoint is emphasized in [8, 9].

This talk draws on joint work with many people, including Bjorn Engquist, Xiantao Li, Di Liu, Weiqing Ren, Eric Vanden-Eijnden and Xingye Yue. It is my pleasure to thank them all for the many helpful discussions that we had, as well as their contributions that lead to the current understanding of HMM. This work is supported in part by ONR and DOE grants.

2 The HMM Framework

The general setting is as follows. We are given a microscopic system whose state variable is denoted by u, together with a microscale model, which can be abstractly written as

$$f(u, b) = 0 \tag{1}$$

where b is the set of auxiliary conditions, such as initial and boundary conditions for the problem. The macroscopic state of the system is denoted by U. U satisfies some abstract macroscopic equation:

$$F(U, D) = 0 \tag{2}$$

where D stands for the macroscopic data that are necessary in order for the model to be complete.

Figure 1: Schematics of the HMM framework

Let us denote by Q the compression or projection operator that maps u to U, and R any operator that reconstructs u from U:

$$Qu = U, \qquad RU = u \tag{3}$$

Q and R should satisfy: $QR = I$ where I is the identity operator. Q is called a compression operator instead of a projection operator since it can be more general than projection. Nevertheless in many cases, it *is* a projection operator. Other terminologies for the same operator include the coarse-graining operator [21], averaging, and restriction [14]. Similarly the reconstruction operator is the standard terminology in shock capturing schemes [16]. They are also called prolongation operator in multi-grid methods, and lifting operator in [14].

HMM consists of two main components.

1. Selection of a macroscopic solver. Even though the macroscopic model is not available completely or is invalid on part of the computational domain, one uses whatever knowledge that is available on the form of F to select a suitable macroscale solver.

2. Estimating the missing macroscale data D using the microscale model. This is typically done in two steps:

   (a) Constrained microscale simulation: At each point where some macroscale data is needed, perform one or a series of constrained microscopic simulations. The microscale solution needs to be constrained so that it is consistent with the local macroscopic state, i.e. $b = b(U)$.

   In practice, this is often the most important technical step.

   (b) Data processing: Use the data generated from the microscopic simulations to extract the needed macroscale data.

Data estimation can either be performed "on the fly" as in a concurrent coupling method, or in a pre-processing step as in a serial coupling method. The latter is often advantageous if the needed data depends on very few variables.

3 The Heterogeneous Multi-scale Finite Element Method

3.1 The statement of the problem

Consider

$$-\nabla \cdot (k^\varepsilon(\mathbf{x})\nabla u^\varepsilon(\mathbf{x})) = f(\mathbf{x}), \qquad \mathbf{x} \in \Omega \subset \mathbb{R}^d. \tag{4}$$

Here ε is a small parameter that signifies explicitly the multi-scale nature of the coefficient $k^\varepsilon(\mathbf{x})$, which will be referred to as the conductivity tensor. Problems of this type have been extensively studied in the context of heat or electric conduction in composite materials, mechanical deformation of composites, etc [3]. In particular, the homogenization technique was initially developed for analyzing these problems [1, 2, 3]. However, except for the case when the microstructure is locally periodic, it is difficult to make use of the homogenized equations for numerical purpose. In addition, the homogenized equation lacks information about the micro-scale behavior which is important for analyzing stress distribution in composites, for example.

In general, we would like to model the macroscopic behavior of the potential field $u^\varepsilon(x)$. In many cases, we would also like to model, at least statistically, the behavior of its gradients.

3.2 The macro-scale solver and the needed data

We will take a finite element approach. For (4), the macro-scale solver can be chosen simply as the standard C^0 piecewise linear finite element method over a macroscopic triangulation $\mathcal{T}_H$ of mesh size H. We will denote by X_H the macroscopic finite element space which could be the standard piecewise linear finite elements over T_H.

The data that need to be estimated from the microscale model is the stiffness matrix on $\mathcal{T}_H$: $A = (A_{ij})$, where

$$A_{ij} = \int_\Omega \nabla\Phi_i(\mathbf{x}) K_H(\mathbf{x}) \nabla\Phi_j(\mathbf{x}) d\mathbf{x}. \tag{5}$$

Here $K_H(\mathbf{x})$ is the effective conductivity tensor at scale H and $\{\Phi_i(\mathbf{x})\}$ are the basis functions for X_H. Had we known $K_H(\mathbf{x})$, we could have evaluated A_{ij} simply by numerical quadrature: Let $f_{ij}(\mathbf{x}) = \nabla\Phi_i(\mathbf{x}) K_H(\mathbf{x}) \nabla\Phi_j(\mathbf{x})$, then

$$A_{ij} = \int_\Omega f_{ij}(\mathbf{x}) d\mathbf{x} \simeq \sum_{T\in\mathcal{T}_H} |T| \sum_{\mathbf{x}_k \in T} \omega_k f_{ij}(\mathbf{x}_k) \tag{6}$$

where $\{\mathbf{x}_k\}$ and $\{\omega_k\}$ are the quadrature points and weights respectively, $|T|$ is the volume of the element $|T|$.

In the absence of explicit knowledge of $K_H(\mathbf{x})$, our problem reduces to the approximation of the values of $\{K_H(\mathbf{x}_k)\}$. This will be done by solving the original microscale model locally around each quadrature point $\{\mathbf{x}_k\}$ (See Figure 2).

Let $I_\delta(\mathbf{x}_k) \ni \mathbf{x}_k$ be a cube of size δ. Consider

$$\nabla \cdot \left(k^\varepsilon(\mathbf{x})\nabla\phi^\varepsilon\right) = 0, \qquad \mathbf{x} \in I_\delta(\mathbf{x}_k). \tag{7}$$

The main objective is to probe efficiently the microscale behavior under the constraint that the average (e.g. macroscale) gradient of the solution ϕ^ε is fixed to be a given constant vector. Having solutions to this local problem, we can define the effective conductivity tensor at $\mathbf{x}_k$ by the relation

$$\langle k^\varepsilon(\mathbf{x})\nabla\phi^\varepsilon\rangle_{I_\delta} = K_H(\mathbf{x}_k)\langle\nabla\phi^\varepsilon\rangle_{I_\delta}, \tag{8}$$

where $\langle v\rangle_{I_\delta} = \frac{1}{|I_\delta|}\int_{I_\delta} v d\mathbf{x}$. The basis of this procedure is the homogenization theorem which has been proved in various contexts; the most general result is found in [17]. The homogenization theorems allow us to define the effective (or homogenized) conductivity tensor, by considering the infinite volume limit of the solutions of the microscale problem subject to the constraint that the average gradient remains fixed. The effective tensor is defined by an average relation of the type (8) in the infinite volume limit, i.e.

$$L = \frac{\delta}{\varepsilon} \to \infty$$

In the special case when the microstructure is periodic, the infinite volume problem reduces to a periodic problem and therefore can be considered on its period.

In practice, one solves (7) with the constraint $\langle\nabla\phi^\varepsilon\rangle_{I_\delta} = \mathbf{e}_1, \cdots, \mathbf{e}_d$ respectively, where d is the spatial dimension of the problem. Denote these solutions by ϕ_j^ε, $j = 1, \cdots, d$. Then

$$(\langle k^\varepsilon(\mathbf{x})\nabla\phi_1^\varepsilon\rangle_{I_\delta}, \cdots, \langle k^\varepsilon(\mathbf{x})\nabla\phi_d^\varepsilon\rangle_{I_\delta}) = K_H(\mathbf{x}_k). \tag{9}$$

In summary, the overall algorithm consists of the following steps:

- Solve for $\phi_1^\varepsilon, \cdots, \phi_\varepsilon^d$ using the boundary conditions discussed below, at each $\mathbf{x}_k$.
- Obtain the approximate values of $K_H(\mathbf{x}_k)$ by averaging the microscale solutions using (9).
- Assemble the effective stiffness matrix using (6).
- Solve the macroscale finite element equation using the effective stiffness matrix. If we express the macroscale solution in X_H in the form of $U_H(\mathbf{x}) = \sum U_j\Phi_j(\mathbf{x})$, then the macroscale finite element equation takes the standard form:

$$AU = F \tag{10}$$

where $U = (U_1, \cdots, U_N)^T, F = (F_1, \cdots, F_N)^T, F_j = (f(\mathbf{x}), \Phi_j(\mathbf{x}))$.

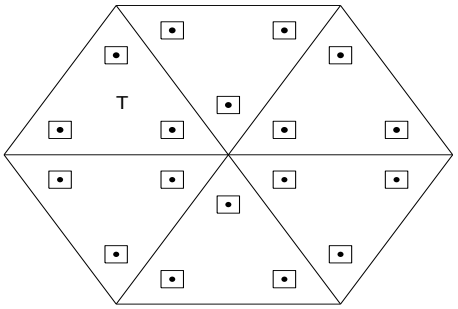

Figure 2: Illustration of HMM for solving (4). The dots are the quadrature points in (6). The little squares are the microcell $I_\delta(\mathbf{x}_k)$.

3.3 The constrained micro-scale solver

The local microscale problem is constrained by the local macroscopic state through the constraint:

$$\langle \nabla \phi^\varepsilon \rangle_{I_\delta} = G \tag{11}$$

for some fixed constant vector G. [24] considered three different types of boundary conditions for the local problem.

1. Dirichlet Formulation.

$$u(\mathbf{x}) = G \cdot \mathbf{x}, \quad \text{on } \partial I_\delta. \tag{12}$$

2. Periodic Formulation.

$$u(\mathbf{x}) - G \cdot \mathbf{x} \text{ is periodic with period } I_\delta. \tag{13}$$

3. Neumann Formulation.

$$k^\varepsilon(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \mathbf{n} = \lambda \cdot \mathbf{n}, \quad \text{on } \partial I_\delta, \tag{14}$$

where the constant vector $\lambda \in \mathbb{R}^d$ is the Lagrange multiplier for the constraint that

$$\langle \nabla u \rangle = G. \tag{15}$$

For example when $d = 2$, to solve problem (14) with the constraint (15), we first solve for u_1 and u_2 from

$$\begin{cases} -\nabla \cdot (k^\varepsilon(\mathbf{x}) \nabla u_i) = 0, & \text{in } I_\delta, \\ k^\varepsilon(\mathbf{x}) \nabla u_i(\mathbf{x}) \cdot \mathbf{n} = \mu_i \cdot \mathbf{n}, & \text{on } \partial I_\delta, \end{cases} \tag{16}$$

for $i = 1, 2$, where $\mu_1 = (1, 0)^T, \mu_2 = (0, 1)^T$. Then given an arbitrary G, the Lagrange multiplier $\lambda = (\lambda_1, \lambda_2)^T$ is determined by the linear equations

$$\lambda_1 \langle \nabla u_1 \rangle + \lambda_2 \langle \nabla u_2 \rangle = G \tag{17}$$

and the solution of (14)-(15) is given by $u = \lambda_1 u_1 + \lambda_2 u_2$.

One can check easily that

$$\langle \nabla u(\mathbf{x}) \rangle \equiv \fint_{I_\delta} \nabla u(\mathbf{x}) d\mathbf{x} \equiv \frac{1}{|I_\delta|} \int_{I_\delta} \nabla u(\mathbf{x}) d\mathbf{x} = G \tag{18}$$

holds for all three formulations.

The performance of these formulations was carefully studied in [24]. The main conclusions were:

1. Periodic boundary condition performs better than the other two formulations.
2. The variance of the estimated effective tensor behaves as $\sigma^2 \sim L^{-d}$ for the random checker-board problem, and $\sigma^2 \sim L^{-2}$ for the periodic problem.
3. In general Neumann formulation underestimates the effective tensor and Dirichlet formulation overestimates the effective tensor. In both cases, the effective conductivity tensors converge to the infinite volume limit with first order accuracy $O(1/L)$, where L is the cell size.

This method was extended in [25] to study the elastic deformation of functionally graded materials.

4 Complex Fluids

4.1 Statement of the problem

Here we discuss how HMM can be applied to model the dynamics of complex fluids to provide "first principle"-based constitutive modeling. This allows us to avoid using empirical constitutive relations that have been popular in traditional approaches. Here "first principle" is in quotation marks since by that we really meant molecular dynamics with empirical potentials. However, the methodology discussed below is also valid if the microscopic model is replaced by first principle-based molecular dynamics.

Using standard notations for liquids, we have the molecular dynamics model as

$$\begin{cases} m\dot{\mathbf{x}}_i(t) = \mathbf{p}_i(t) \\ \dot{\mathbf{p}}_i(t) = \mathbf{F}_i \end{cases} \tag{19}$$

$i = 1, 2, \cdots, N$. Here $\mathbf{x}_i$ and $\mathbf{p}_i$ are its position and momentum respectively, $\mathbf{F}_i$ is the force acting on the i-th particle. We will work in the isothermal setting, and the Nosé-Hoover thermostat can be used to control the temperature of the system.

We will assume that the atomic forces are short ranged. More complex atomistic models are often required to model complex fluids accurately. For example, one might

use a bead-spring model to represent flexible polymers. Here we will not consider such complications, since the basic methodolgy remains the same.

At the continuum level, the dynamics of incompressible flow have to obey conservation laws of mass and momentum:

$$\begin{cases} \rho \partial_t \mathbf{u} = \nabla \cdot \tau \\ \nabla \cdot \mathbf{u} = 0 \end{cases} \tag{20}$$

where the momentum flux $-\tau = \rho\, \mathbf{u} \otimes \mathbf{u} - \tau_d$. Here ρ is the density of the fluid which is assumed to be a constant, $\mathbf{u} = (u, v)$ is the velocity field, and τ_d is the stress tensor. At this stage the system is not closed since the stress tensor is yet to be specified. Traditionally the idea has been to close this system by an empirically postulated constitutive relation, such as

$$\tau_d = -pI + \mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) \tag{21}$$

for simple fluids. Here we will develop numerical methods that bypass such empirical constitutive modeling.

Another important component in the model is the boundary condition. Here we will assume the standard no-slip boundary condition

$$\mathbf{u} = \mathbf{u}_0 \tag{22}$$

where $\mathbf{u}_0$ is the velocity of the boundary. The issue of boundary condition is another important source of problems for multi-scale modeling, see [20]. But here we will skip this issue.

4.2 The macro-scale solver and the needed data

As the macroscopic solver, we choose the projection method on a staggered grid [6]. Projection method is a fractional step method. At each time step, we first discretize the time derivative in the momentum equation by the forward Euler scheme to obtain an intermediate velocity field:

$$\rho \frac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} = \nabla \cdot \tau^n \tag{23}$$

where τ^n is the momentum flux. For the moment, pressure as well as the incompressibility condition are neglected. Next the velocity field $\tilde{\mathbf{u}}^{n+1}$ is projected onto the divergence-free subspace:

$$\rho \frac{\mathbf{u}^{n+1} - \tilde{\mathbf{u}}^{n+1}}{\Delta t} + \nabla p^{n+1} = 0 \tag{24}$$

where p^{n+1} is determined by

$$\Delta p^{n+1} = \frac{\rho}{\Delta t} \nabla \cdot \tilde{\mathbf{u}}^{n+1} \tag{25}$$

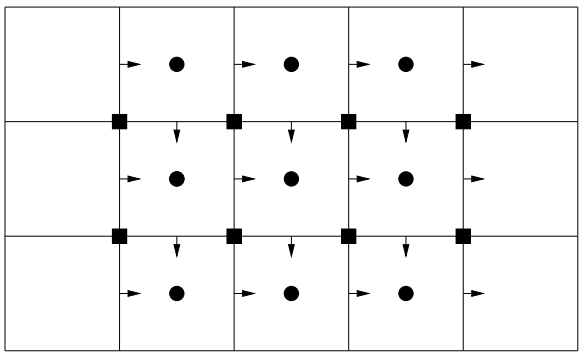

Figure 3: Schematic of the spatial discretization of the continuum equations in (20). u is defined at $(x_i, y_{j+\frac{1}{2}})$, v is defined at $(x_{i+\frac{1}{2}}, y_j)$, and p is at the cell center $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$. τ_{11} and τ_{22} are calculated at the cell center indicated by circles, and τ_{12} is calculated at the grid points indicated by squares.

usually with Neumann boundary condition.

The spatial discretization is shown in Figure 3. For integer values of i and j, we define u at $(x_i, y_{j+\frac{1}{2}})$, v at $(x_{i+\frac{1}{2}}, y_j)$, and p at the cell center $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$. The diagonals of the flux τ are defined at $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$, and the off-diagonals are defined at (x_i, y_j). The operators ∇ and Δ are discretized by standard central difference and the five-point formula respectively. The use of this grid simplifies the coupling with molecular dynamics.

The data that need to be estimated from molecular dynamics are the stresses. Here we will make a constitutive assumption, namely that the stress depends only on the rate of strain. We do not need to know anything about the specific functional form of this dependence.

4.3 The constrained micro-scale solver: Constant rate-of-strain molecular dynamics

The key component in estimating the stress is to construct a constant rate-of-strain ensemble for the MD. This is done through a modified periodic boundary condition [20]. The main idea is to perform molecular dynamics using periodic boundary condition, by the period itself is changing in time in such a way that its vertices solve the ODE:

$$\dot{\mathbf{x}} = A\mathbf{x}, \tag{26}$$

where A is the rate of strain tensor. In the special case when A represents a pure shear

$$A = \begin{pmatrix} 0 & b & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{27}$$

this boundary condition becomes simply the Lees-Edwards boundary condition [15]: A constant shear profile is maintained by shifting the periodic copies of the simulation box above and below in opposite directions according.

If this idea is used naively, one encounters the difficulty that the period box might be severely deformed. This is easy to see in the special case of a extensional flow, in which one side of the period box grows exponentially in time. A clever reinitialization procedure has been developed to handle this difficulty [20].

The MD simulation keeps track of the positions and velocities of all particles as functions of time, from which the momentum flux tensor can be calculated using the Irving-Kirkwood formula (28):

$$\begin{aligned}\tau(\mathbf{x},t) &= -\sum_i \tfrac{1}{m_i}(\mathbf{p}_i(t)\otimes\mathbf{p}_i(t))\delta(\mathbf{x}_i(t)-\mathbf{x}) \\ &\quad -\tfrac{1}{2}\sum_{j\neq i}\left((\mathbf{x}_i(t)-\mathbf{x}_j(t))\otimes\mathbf{F}_{ij}(t)\right)\int_0^1\delta(\lambda\mathbf{x}_i(t)+(1-\lambda)\mathbf{x}_j(t)-\mathbf{x})\,d\lambda\end{aligned} \tag{28}$$

where $\mathbf{F}_{ij}(t)$ is the force acting on the i-th particle by the j-th particle. When (28) is averaged over the simulation box, it gives

$$\tau(t) = -\frac{1}{|\Omega|}\sum_{\mathbf{x}_i\in\Omega(t)}\frac{1}{m_i}(\mathbf{p}_i\otimes\mathbf{p}_i) - \frac{1}{2|\Omega|}\sum_{j\neq i} d_{ij}(\mathbf{x}_i-\mathbf{x}_j)\otimes\mathbf{F}_{ij} \tag{29}$$

where in the first term the summation runs over particles inside the box, and in the second term the summation is over all pairs of particles including their images. Here d_{ij} is defined as

$$d_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i,\mathbf{x}_j\in\Omega, \\ 0, & \text{if } \mathbf{x}_i,\mathbf{x}_j\notin\Omega, \\ c, & \text{if only one of } \mathbf{x}_i,\mathbf{x}_j \text{ in } \Omega \end{cases} \tag{30}$$

where $0\le c\le 1$ is the fraction of $|\mathbf{x}_i-\mathbf{x}_j|$ being cut by the box. Therefore besides contributions from particles inside the box, particles outside the box also contribute to the stress.

(29) gives the instantaneous stress at the microscopic time t. To extract the macroscopic stress, (29) is averaged over time to give an estimate for the macroscale stress:

$$\tau = \frac{1}{T-T_0}\int_{T_0}^{T}\tau(t)dt \tag{31}$$

where T_0 is some relaxation time.

The overall algorithm looks as follows at each macro time step:

1. Step 1. Calculate the needed stresses by constrained local MD simulations;
2. Step 2. Using the projection method and the computed stresses to get $\mathbf{u}^{k+1}$.

For numerical results computed using this algorithm, we refer to [20, 19].

5 Chemical Kinetic Systems with Multiple Time Scales

5.1 Statement of the problem

This last example is selected to illustrate issues for problems with multiple time scales. We will discuss chemical kinetic systems modeled by discrete Markov chains. The standard approach for modeling such systems is the stochastic simulation algorithm (SSA), also called the Gillespie algorithm. Since the rate functions tyically have exponential dependence on the activation energy, it is common for such systems to exhibit many different time scales. These problems have received a great deal of attention in recent years, and we refer to [10] for references and discussions of recent work on this problem. Here for illustration, we will focus on the simplest situation of problems with two separated time scales.

We begin with the general set-up. Assume that an isothermal system has N_S species of molecules S_i, $i = 1, \ldots, N_S$, and there are M_R reaction channels R_j, $j = 1, \ldots, M_R$. Let x_i be the number of molecules of species S_i. Then the state of the system is given by the vector

$$x = (x_1, \cdots, x_{N_R}) \tag{32}$$

Each reaction R_j is characterized by a rate function $a_j(x)$ and a vector ν_j that describes the change of the state due to the reaction. We write

$$R_j = (a_j, \nu_j). \tag{33}$$

The dynamics of the model is completely specified by the following rules:

1. Given the state x, the reactions are independent of each other on an infinitesimal time interval of duration dt and the probability for the reaction R_j to happen is given by $a_j(x)dt$.

2. The state of the system after R_j is given by $x + \nu_j$.

The standard computer implementation of such a model is given by the well-known stochastic simulation algorithm (SSA) proposed in [13]. Let

$$a_0(x) = \sum_{j=1}^{M_R} a_j(x). \tag{34}$$

Assume that the current time is $t = t_n$, and the state of the system is $x = x_n$. The essential steps of SSA are the following:

1. Generate independent random numbers r_1 and r_2 with uniform distribution on the unit interval. Let
$$\delta t_{n+1} = -\frac{\ln r_1}{a_0(x)}, \tag{35}$$
and k_{n+1} to be the natural number such that
$$\frac{1}{a_0(x)} \sum_{j=1}^{k_{n+1}-1} a_j(x) < r_2 \leq \frac{1}{a_0(x)} \sum_{j=1}^{k_{n+1}} a_j(x). \tag{36}$$

2. Update the time and the state of the system
$$t_{n+1} = t_n + \delta t_{n+1}\ , \qquad x_{n+1} = x_n + \nu_{k_{n+1}}. \tag{37}$$

Next we consider the case when the rates are divided into two groups (the general case with more than two groups is treated in [10, 11]): One group corresponding to the fast processes with rates of order $1/\varepsilon$ and one group corresponding to the slow processes with rates of order 1. Here $\varepsilon \ll 1$:
$$a(x) = \big(a^s(x), a^f(x)\big), \tag{38}$$
where
$$\begin{aligned} a^s(x) &= \big(a_1^s(x), \cdots, a_{M_s}^s(x)\big) = O(1), \\ a^f(x) &= \big(a_1^f(x), \cdots, a_{M_f}^f(x)\big) = O(1/\varepsilon). \end{aligned} \tag{39}$$
in dimensionless units. The corresponding reactions and the associated state change vectors can be grouped accordingly:
$$R^s = (a^s, \nu^s), \qquad R^f = (a^f, \nu^f). \tag{40}$$
As a simple example, consider the system:
$$S_1 \underset{a_2^f}{\overset{a_1^f}{\rightleftarrows}} S_2, \qquad S_2 \underset{a_2^s}{\overset{a_1^s}{\rightleftarrows}} S_3, \qquad S_3 \underset{a_4^f}{\overset{a_3^f}{\rightleftarrows}} S_4, \tag{41}$$
with reaction channels given by
$$\begin{aligned} a_1^f &= 10^5 x_1, & \nu_1^f &= (-1, 1, 0, 0); \\ a_2^f &= 10^5 x_2, & \nu_2^f &= (1, -1, 0, 0); \\ a_3^f &= 10^5 x_3, & \nu_3^f &= (0, 0, -1, 1); \\ a_4^f &= 10^5 x_4, & \nu_4^f &= (0, 0, 1, -1); \\ a_1^s &= x_2, & \nu_1^s &= (0, -1, 1, 0); \\ a_2^s &= x_3, & \nu_2^s &= (0, 1, -1, 0). \end{aligned} \tag{42}$$
There are a total of 4 species and 6 reaction channels, with 4 fast reactions and 2 slow ones. We can think of ε as being 10^{-5} for this example.

5.2 The macro-scale solver and the data to be estimated

For problems of this type, direct application of SSA will result in time steps of size ε (the total rate $a_0(x)$ in (34) is of order $1/\varepsilon$) with a total cost of order $1/\varepsilon$ if we want to advance the whole system through a time interval of order unity. Most of the cost will be spent on the fast reactions, which are often of little interest in these cases. Indeed for such systems, we are usually interested in the slow processes since they are the rate-limiting steps. Here we propose a modified SSA that captures the slow processes at a cost that is independent of ε, and therefore much less than that of the direct SSA when $\varepsilon \ll 1$. The underlying assumption is that the fast processes are in equilibrium over the slow time scale.

The macro-scale solver is a SSA for the slow reactions only. The data that need to be estimated are the effective rates, given that the fast processes are in quasi-equilibrium.

5.3 The micro-scale solver and the overall algorithm

The effective rates for slow reactions are estimated using another SSA or several independent replicas of the SSA, but using only the fast reactions. The overall algorithm takes the form of a nestd SSA: The outer SSA is on the slow processes only, but uses modified slow rates. The inner SSA is on the fast processes only: it uses the original fast rates and serves to give the modified slow rates. Let $t = t_n$, $x = x_n$ be the current time and state of the system respectively. Given (t_n, x_n), do:

1. **Inner SSA**
   Run N independent replicas of SSA with the fast reactions $R^f = (a^f, \nu^f)$ only, for a time interval of T_f. During this calculation, compute the modified slow rates: For $j = 1, \cdots, M_s$, these are

$$\tilde{a}_j^s = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{T_f} \int_{T_0}^{T_f+T_0} a_j^s(x_k(\tau)) d\tau, \tag{43}$$

   where $x_k(\tau)$ is the k-th replica of the auxiliary fast process at virtual time τ whose initial value is $x_k(0) = x_n$, T_0 is a parameter we choose in order to minimize the effect of the transients in the auxiliary fast process.

2. **Outer SSA**
   Run one step of SSA for the modified slow reactions $\tilde{R}^s = (\tilde{a}^s, \nu^s)$ to generate (t_{n+1}, x_{n+1}) from (t_n, x_n).

Then repeat.

Note that the number of replicas N can be taken to be one in the above algorithm, but it is advantageous to take a large N since the Inner SSA can be trivially parallelized.

Examples of results using this nested SSA algorithm and comparison with the direct SSA algorithm can be found in [10, 11].

This is an example of the seamless HMM, i.e. no specification of slow variables is needed and no explicit constraints are applied on the inner SSA. We should remark that there are very few examples for which seamless HMM works. The reasons behind the present example is analyzed in [11].

5.4 Slow variables and effective dynamics

Even though the formulation of the nested SSA does not require explicit identification of the slow variables and the effective dynamics over the slow time scale, to understand why and how the algorithm works, we do need to understand these issues.

First we discuss the observables, which are functions of the state variable x. By definition, slow observables are conserved quantities during the fast reactions, i.e. $v(x)$ is a slow observable if for any $x \in X$ and any state change vector ν^f associated with the fast reactions one has

$$v(x + \nu^f) = v(x). \tag{44}$$

This means that the value of the slow observable $v(x)$ is unaffected by the fast reactions. To find a general representation of such observables, we consider special slow observables which are linear functions that satisfy (44). We call such slow observables slow variables. It is easy to see that $v(x) = b \cdot x$ is a slow variable if

$$b \cdot \nu^f = 0, \tag{45}$$

for all ν^f. Let $b_1, b_2, \ldots, b_J$ be a set of basis vectors that satisfy (45). Define

$$y_j = b_j \cdot x \quad \text{for} \quad j = 1, \ldots, J. \tag{46}$$

Then $y_1, y_2, \cdots, y_J$ defines a complete set of slow variables, i.e. all slow observables can be expressed as functions of $y_1, y_2, \cdots, y_J$. For the example considered earlier, it is easy to see that both $x_1 + x_2$ and $x_3 + x_4$ are conserved during the fast reactions, i.e. $y_1 = x_1 + x_2$ and $y_2 = x_3 + x_4$ are the slow variables of that system.

We can now put the quasi-equilibrium assumption in precise terms. Fix the slow variables y and consider the virtual fast process defined by retaining only the fast reaction channels [4]. Two important conclusions can be drawn for this process. The first is that the slow variables are held constant. The second is that this system approaches a unique equilibrium state (which depends on the value of y) on a time-scale of order ε. This equilibrium state is the desired quasi-equilibrium, which we denote by $\mu_y(x)$. The rates for the effective slow process are obtained by averaging the slow rates with respect to this quasi-equilibrium:

$$\bar{a}_j(y) = \langle a_j(x) \rangle_y \equiv \sum_{x \in X} a_j(x) \mu_y(x). \tag{47}$$

It is obvious that the effective rates are only functions of y. The effective dynamics is completely specified by

$$\bar{R} = (R^s, \bar{a}(y)) \ . \tag{48}$$

It is shown in [11] by singular perturbation analysis that the original dynamics converges to the above effective dynamics with an error of order $O(\varepsilon)$.

5.5 Convergence and efficiency of the nested SSA

If we knew $\bar{a}(y) = (\bar{a}_1(y), \cdots, \bar{a}_{M_s}(y))$ explicitly, we could have carried out SSA using these rates. This would capture the process on the slow time-scale, which is what we are interested in. For convenience, we will call such a procedure "averaged SSA". Unfortunately we usually do not have an explicit expression for the effective rates (47). The nested SSA proposed above is a way of getting approximate values of these rates "on-the-fly".

To see why this algorithm should work, it is clear that the only difference between the nested SSA and the averaged SSA is that the averaged SSA uses the rates in (47), whereas nested SSA uses the rates in (43). However, by ergodicity we have that $\tilde{a}$ converges to $\bar{a}$ for when T_f and N goes to infinity. Therefore the results of the two algorithms also become closer and closer for large T_f and N. Quantitative error estimates can be obtained. The details are given in [11]. Among other things, it is proved in [11] that

$$\mathbb{E}|\bar{a}_j - \tilde{a}_j| \le C\Big(\frac{e^{-\alpha T_0/\varepsilon}}{1+T_f/\varepsilon} + \frac{1}{\sqrt{N(1+T_f/\varepsilon)}}\Big), \tag{49}$$

for some constants C and α which are independent of ε. Here $\mathbb{E}$ denotes expectation with respect to the statistics of the virtual fast process in the Inner SSA. The first term on the right hand-side of (49) measures the deviation from the quasi-equilibrium if the inner SSA is run only for a time duration of T_f, starting at T_0. α measures the rate of convergence for the virtual fast process to equilibrium. The second term measures the sampling error from using time and ensemble averaging on a time interval of duration T_f with an ensemble of N replicas.

Let us now estimate the cost of the nested SSA. For simplicity we will take $T_0 = 0$ here and also in our numerical examples. One feature of (49) is that this estimate depends on the ratio T_f/ε rather than T_f alone. This means that, when $\varepsilon \ll 1$, we can achieve a small error on $\tilde{a}_j$ by choosing $T_f/\varepsilon \gg 1$ and yet have $T_f \ll 1$ (remember that we have assumed that the time-scale for the slow process is of order 1). This is the very reason why the nested SSA is more efficient than a direct SSA. To quantify this, suppose that we want to compute the results within an error tolerance λ. To control each term in (49) by λ, the optimal choice of parameters is:

$$N = T_f/\varepsilon = 1/\lambda^2. \tag{50}$$

Then the cost for evolving the nested SSA for a unit time is estimated to be:

$$\text{cost} = \tilde{a}_0^s \times N a_0^f T_f/\epsilon = O\big(1/\lambda^2\big), \tag{51}$$

which is independent of ε.

6 Conclusions

Let us now summarize.

6.1 Main features of HMM

HMM is a top-down framework: It is based on the macroscopic solver and uses the microscale model as a supplement. Its two-component procedure offers a number of interesting features, including:

1. The ability to make maximum use of our knowledge on all scales. For example, when choosing the macroscale solver, we may take advantage of what is known at the macroscale, such as conservation form, variational structure, etc, as well as the nature of the physical process such as shock formation, phase transformation.

2. The flexibility to make maximum use of the special features of the problem. Time scale separation, for example, is made use of by the fact that satisfactory approximation for the needed data can be obtained by carrying out the microscale solver on relatively short time intervals.

   The latter is really the key to multi-scale modeling: Our aim is not to construct general purpose methods, but rather to construct particularly efficient methods by taking advantage of the special features of the problem.

6.2 Other general methodologies

Domain decomposition and adaptive model refinement are two popular and general methodologies that have been explored for problems where microscopic models are needed in the vicinity of localized defects, such as cracks or contact lines.

"Equation-free" has been proposed as another general methodology for first-principle-based constitutive modeling [14]. The philosophy of "equation-free" is the same as that of concurrent coupling, namely the needed constitutive information is not computed beforehand, but rather "on-the-fly" as the computation proceeds. The technical details of "equation-free", which are embodied in the "gap-tooth schemes", "patch dynamics", etc, are not sufficient at this point to allow us to discuss their merits when applied to realistic problems. In the simplest cases such as stiff ODEs, "equation-free" becomes quite similar to HMM. But for more complicated situations such as

stochastic ODEs, oscillatory ODEs, or for problems with spatial microstructures, it is unclear how "equation-free" would proceed in order obtain acceptable accuracy and at the same time, beat the brute force microscopic solvers in terms of efficiency. See [9].

6.3 The fiber bundle viewpoint

Finally, it is worth emphasizing that the conceptual framework of HMM is very similar to that of a fiber bundle. It is true that for most problems, both the microscopic and the macroscopic variables are defined on the same physical space. Domain decomposition and adaptive model refinement methods rely on this fact. For HMM, it is helpful to think about the local microstructure as if it is defined on a different space, the fibers over the macroscopic space. Indeed, under the framework of HMM, the microscale computations carried out over the different macroscale locations do not communicate with each, except through the macroscale solver. This is very much reminiscent of the structure of fiber bundles. By thinking of the microstructure as been defined in a virtual space, we are naturally led to numerical algorithms which are free of the limitations associated with filling up macroscale space and time by the microscale grid points and time steps, and this is the reason why HMM enables us to design numerical algorithms that are much more efficient than the brute force microscale solvers.

References

[1] I. Babuska, "Homogenization and Its Applications, Mathematical and Computational Problems," in *Numerical Solutions of Partial Differential Equations-III*, (SYNSPADE 1975, Maryland, May 1975), B. Hubbard (ed.), Academic Press, New York, pp. 89–116, 1976.

[2] I. Babuska, "Solution of Interface by Homogenization, I, II, III," *SIAM J. Math. Anal.*, vol. 7, pp. 603–645, 1976; vol. 8, pp. 923–937, 1977.

[3] A. Bensoussan, J.L. Lions and G.C. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland Pub. Co., Amsterdam, 1978.

[4] Y. Cao, D. Gillespie, and L. Petzold, preprint.

[5] R. Car and M. Parrinello, "Unified Approach for Molecular Dynamics and Density-Functional Theory," *Phys. Rev. Lett.*, vol. 55, no. 22, pp. 2471–2474, 1985.

[6] A. J. Chorin, "Numerical Solution of the Navier-Stokes Equations," *Math. Comp.*, vol. 22, pp. 745–762, 1968.

[7] W. E and B. Engquist, "The Heterogeneous Multi-Scale Methods," *Comm. Math. Sci.*, vol. 1, pp. 87–133, 2003.

[8] W. E and B. Engquist, "Multiscale Modeling and Computation," *Notices of the AMS*, vol. 50, no. 9, pp. 1062-1070, 2003.

[9] W. E, B. Engquist, X. Li, W. Ren and E. Vanden-Eijnden, "The Heterogeneous Multi-scale Method: A Review", preprint.

[10] W. E, D. Liu and E. Vanden-Eijnden, "Nested Stochastic Simulation Algorithm for Chemical Kinetic Systems with Disparate Rates", to appear in J. Chem. Phys.

[11] W. E, D. Liu and E. Vanden-Eijnden, "Nested Stochastic Simulation Algorithm for Multi-scale Systems", preprint.

[12] C.W. Gear and I.G. Kevrekidis, "Projective Methods for Stiff Differential Equations: Problems with Gaps in Their Eigenvalue Spectrum," *SIAM J. Sci. Comput.*, vol. 24, no. 4, pp. 1091–1106, 2003.

[13] D. T. Gillespie, *J. Comp. Phys.* **22**, 403–434 (1976).

[14] I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg and C. Theodoropoulos, "Equation-Free, Coarse-Grained Multiscale Computation: Enabling Microscopic Simulators to Perform System-Level Analysis," *Comm. Math. Sci.*, vol. 1, no. 4, pp. 715–762, 2003.

[15] A.W. Lees and S.F. Edwards, "The Computer Study of Transport Processes Under Extreme Conditions," *J. Phys. C*, vol. 5, no. 15, pp. 1921–1929, 1972.

[16] R. LeVeque, *Numerical Methods for Conservation Laws*, Birkhauser, 1990.

[17] G.C. Papanicolaou and S.R.S. Varadhan, "Boundary Value Problems with Rapidly Oscillating Random Coefficients," in *Random Fields*, J. Fritz, et.al (eds.), vol. 27, Amsterdam: North-Holland, pp. 835–873, 1981.

[18] B. Perthame and F. Coron, "Numerical Passage from Kinetic to Fluid Equations," *SIAM J. Numer. Anal.*, vol. 28, pp. 26–42, 1991.

[19] W. Ren, in preparation.

[20] W. Ren and W. E, "Heterogeneous Multiscale Method for the Modeling of Complex Fluids and Microfluidics", *J. Comput. Phys.*, 2005.

[21] R.E. Rudd and J.Q. Broughton, "Coarse-Grained Molecular Dynamics and the Atomic Limit of Finite Elements," *Phys. Rev. B*, vol. 58, no. 10, pp. R5893–R5896, 1998.

[22] E.B. Tadmor, M. Ortiz and R. Phillips, "Quasicontinuum Analysis of Defects in Crystals," *Philos. Mag.*, A73 , 1529–1563, 1996.

[23] E. Vanden-Eijnden, "Numerical Techniques for Multiscale Dynamical Systems with Stochastic Effects,' *Comm. Math. Sci.*, vol. 1, no. 2, pp. 385-391, 2003.

[24] X.-Y. Yue and W. E, " The Local Microscale Problem in the Multiscale Modelling of Strongly Heterogeneous Media: Effect of Boundary Conditions and Cell Size", submitted to J. Comput. Phys.

[25] X.-Y. Yue and W. E, "A Numerical Method for Studying the Properties of Functionally Graded Composite Materials", in preparation.