Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2018-0174

Vol. **25**, No. 5, pp. 1259-1301
May 2019

Physics-Constrained, Data-Driven Discovery of Coarse-Grained Dynamics

Lukas Felsberger[1] and Phaedon-Stelios Koutsourelakis[1,*]

[1] *Continuum Mechanics Group, Department of Mechanical Engineering, Technical University of Munich, Boltzmannstrasse 15 85748, Garching, Germany.*

Received 12 February 2018; Accepted (in revised version) 6 July 2018

Abstract. The combination of high-dimensionality and disparity of time scales encountered in many problems in computational physics has motivated the development of coarse-grained (CG) models. In this paper, we advocate the paradigm of data-driven discovery for extracting governing equations by employing fine-scale simulation data. In particular, we cast the coarse-graining process under a probabilistic state-space model where the transition law dictates the evolution of the CG state variables and the emission law the coarse-to-fine map. The directed probabilistic graphical model implied, suggests that given values for the fine-grained (FG) variables, probabilistic inference tools must be employed to identify the corresponding values for the CG states and to that end, we employ Stochastic Variational Inference. We advocate a sparse Bayesian learning perspective which avoids overfitting and reveals the most salient features in the CG evolution law. The formulation adopted enables the quantification of a crucial, and often neglected, component in the CG process, i.e. the predictive uncertainty due to information loss. Furthermore, it is capable of reconstructing the evolution of the full, fine-scale system. We demonstrate the efficacy of the proposed framework in high-dimensional systems of random walkers.

AMS subject classifications: 62F15, 82C21, 82C80

Key words: Coarse-graining, dynamics, Bayesian, non-equilibrium, data-driven.

1 Introduction

The present paper is concerned with the discovery of data-driven, dynamic, stochastic coarse-grained models from fine-scale simulations with a view of advancing multiscale modeling. Many problems in science and engineering are modeled by high-dimensional systems of deterministic or stochastic, (non)linear, microscopic evolution laws (e.g. ODEs). Several such examples are encountered in quantum mechanical models [1], in the atomistic simulation of materials [2], in complex flows [3–5], and in agent-based models [6].

*Corresponding author. *Email addresses:* p.s.koutsourelakis@tum.de (P. S. Koutsourelakis), lukas.felsberger@cern.ch (L. Felsberger)

Their solution is generally dominated by the smaller time scales involved even though the outputs of interest might pertain to time scales that are greater by several orders of magnitude [7]. The combination of high-dimensionality and disparity of time scales has motivated the development of coarse-grained formulations. These aim at constructing a much lower-dimensional model that is practical to integrate in time and can adequately predict the outputs of interest over the time scales of interest. The literature on this topic is vast and many categorizations are possible. This paper focuses on *data-driven* strategies [8–18] where (short) simulations of the original, fine-grained (FG) or full-order system of equations are used in order to learn (or infer) the right coarse-grained (CG) model. This is consistent with the emergence of data-driven discovery, commonly referred to as the fourth paradigm in science [19]. Extracting governing equations from data is a central challenge in a wide variety of physical and engineering sciences as in climate science, neuroscience, ecology, finance, and epidemiology, where models (or closures) often remain elusive [20–23].

The challenge in multiscale physical systems such as those encountered in non-equilibrium statistical mechanics [24], is even greater as apart from the identification of an effective model, it is crucial to discover a good set of CG state variables [25]. The latter problem has been traditionally addressed separately from the construction of the CG model, using dimensionality reduction techniques, (e.g. [26]) but in this work we offer a holistic treatment.

The starting point of all (to the best of our knowledge) CG schemes is the specification of a set of coarse (or reduced) state variables with respect to which a CG model is prescribed/found. These variables are selected on the basis of the analysis objectives, or because they are known to be "slow" either through physical insight or using rigorous statistical learning tools (e.g. dimensionality reduction). The specification of these coarse variables is generally done through a projection (or restriction), many-to-one operator from the fine-scale variables to the coarse. For problems exhibiting time-scale separation and when the CG state variables identified correspond to slowly-evolving features, it has been established that Markovian CG models are suitable and several numerical strategies have proven successful [27, 28]. In the absence of scale separation however, memory effects (dissipation) and thermal noise play an important role. The Mori-Zwanzig (MZ) formalism which was originally developed in the context of irreversible statistical mechanics [29, 30] but has been extended to general systems of ODEs (e.g. [31–34]) provides a rigorous mathematical foundation. Setting aside the largely unsolved computational challenges associated with finding the terms in the Generalized Langevin Equation (GLE) [35, 36] prescribed by MZ, it is in principle a perfect scheme, i.e. it is capable of capturing exactly the dynamics of the state variables that are resolved and with respect to which a closed system of equations is written. The solution of non-Markovian system of equations (due to the memory term) poses itself several difficulties as it necessitates storing the history of the CG variables. Interestingly, in many instances, this non-Markovian system can be mapped onto an, albeit higher-dimensional, Markovian, system with additional degrees of freedom corresponding to auxiliary or extended state

variables. Such a strategy has been employed in other models exhibiting memory, e.g. in plasticity of solids where the auxiliary variables are physically motivated (internal state variables, [37]). Another difficulty stems from sampling of the random noise term in the GLE which is essential in achieving the correct equilibrium statistics according to the second fluctuation-dissipation theorem (FDT) [38].

Even if a good CG model can be identified such that it correctly predicts the time evolution of these coarse variables, it is difficult (without additional assumptions such as the prescription of a lifting operator, [39]) to reconstruct or infer the evolution of the full, fine-scale system. Hence if the observables one is interested in predicting, do not exclusively depend on the coarse variables selected, the CG model constructed is not useful. In this paper, we advocate a different strategy, i.e. we aim at constructing a reduced or coarsened description of the FG system that is predictive of the whole fine-scale picture. The starting point is the prescription of a (parametrized) *probabilistic*, coarse-to-fine map [14]. As a result, the CG state variables are defined indirectly and are latent (hidden) given the FG simulation data. The formulation is complemented by a (parametrized) *probabilistic* evolution law for the CG variables. The goal is not to write a consistent evolution law with respect to these state variables (as one would do with MZ when possible), but rather to identify the right evolution law for the CG states that is most predictive of the FG futures.

The formulation advocated enables the quantification of a crucial, and often neglected, component in the CG process [40], i.e. the predictive uncertainty due to information loss that unavoidably takes place. That is, unless there are known redundancies in the full-order, FG description, the coarse variables will provide a lossy compression. This in turn will manifest itself in uncertainty in the predictions that we can generate given the coarse variables (i.e. there are many microscopic states compatible with each value of the CG variables). We emphasize that this has nothing to do with the quality of the coarse model nor with its potentially stochastic nature. Even if the exact equations were available (e.g. through MZ) i.e. one was able to perfectly predict the time evolution of the coarse state variables, it is not necessary that one would be able to predict perfectly the full, fine-scale picture. Naturally, if the CG model was stochastic, the randomness in the evolution of the CG states would compound the aforementioned predictive uncertainty.

Another potentially significant source of uncertainty in the context of data-driven techniques, pertains to the data itself. That is, the CG model (or the parameters thereof) apart from being data-dependent, is learned from finite amounts of fine-scale simulation data. As these are generally expensive to procure and only relatively small time spans are accessible by microscopic simulators (e.g. molecular dynamics), one would like to minimize the number necessary or even better, to be able to quantify the effect of a finite-sized dataset to the uncertainty in the predictive estimates. In this manner the analyst can decide (potentially also from an optimal experimental design perspective [41]) whether the predictive accuracy of the CG model is adequate or additional FG simulation data is needed.

In this paper, we treat the CG model as a probabilistic state-space model [42–44]

where the transition law dictates the evolution of the CG state variables and the emission law the coarse-to-fine map. The directed probabilistic graphical model implied, suggests that given values for the FG variables, inference tools must be employed to identify the corresponding values for the CG states [45]. Naturally, one of the most critical questions pertains to the form of the CG evolution law which poses a formidable model selection issue. Correct identification of the right-hand side terms involved can also reveal qualitative features of the coarse-scale evolution such as the type of constitutive relations [20]. It is crucial therefore to be able to search across model types. In order to avoid overfitting as well as endow the estimates with robustness even in the presence of limited data, we invoke the principle of parsimony that is reflected in the *sparsity* of the solutions. Sparsity, as a guiding principle in the identification of dynamical systems has been employed before in [46–48]. Naturally the topic has a very long history in image/signal processing with far reaching contributions in the context of unsupervised learning [49–53] which most closely resembles the framework used here. We adopt a sparse Bayesian learning [54,55] perspective whereby appropriate hyperpriors based on Automatic Relevance Determination (ARD, [56]) and without the specification of any additional hyperparameters, can achieve this goal i.e. regularize the solution space using a data-dependent prior distribution that effectively removes redundant or superfluous features. The model advocated is combined with Stochastic Variational Inference procedures for the latent CG states [57,58].

We emphasize that the observations/data employed are not direct i.e. they do not explicitly correspond to the CG state variables as in [22, 59, 60]. This would effectively require solving a regression problem [61]. Instead, the CG state variables are actually inferred, simultaneously with the CG model form/parameters. As a result they can subsequently be used to reconstruct FG state variables futures either for the purposes of computing various observables or simply for consistently re-initializing the data-generating FG simulator.

The structure of the rest of the paper is as follows. In section 2 we present the general methodological framework and introduce the probabilistic graphical model underlying the CG process. We provide a specification of the advocated for systems of random walkers (agents) and discuss the issues of enforcement of conservation laws as well model-structure discovery. We conclude this section with algorithmic details pertaining to the inference (training) and prediction steps. In section 3 we demonstrate the capabilities of the proposed formulation in three examples involving high-dimensional systems of random walkers.

2 Methodology

This section introduces the notational conventions adopted and presents the proposed modeling and computational framework. Subscripts with "f" denote quantities pertaining to the FG (or fine-scale or microscopic) description whereas those with "c" to the CG

(or coarse-scale or macroscopic) model. We also generally employ lowercase letters (e.g. $\boldsymbol{x}$) to denote quantities related to the FG model and uppercase (e.g. $\boldsymbol{X}$) for CG-model-related ones. Generally superscripts in parentheses e.g. $\boldsymbol{x}^{(i)}$ or $\boldsymbol{x}^{(I)}$ indicate sample numbers.

2.1 General probabilistic state-space model

We start with the presentation of the basic components of the proposed state-space model in a rather general setting. One of the primary goals is to demonstrate the flexibility of the data-driven strategy advocated, the Bayesian nature of the learning process as well as its suitability in producing *probabilistic* predictions that quantify the information loss and reflect the finiteness of data.

We consider a (generally high-dimensional) FG model with state variables $\boldsymbol{x}\in\mathcal{M}_f\subset\mathbb{R}^{n_f}$ (where $n_f\gg 1$), the dynamics of which are described by a system of deterministic or stochastic ODEs, i.e.

$$\begin{cases} \dot{\boldsymbol{x}}_t=\boldsymbol{f}(\boldsymbol{x}_t), \\ \boldsymbol{x}_0=\hat{\boldsymbol{x}}_0 \text{ or } \boldsymbol{x}_0 \text{ drawn from } p_{f,0}(\boldsymbol{x}_0), \end{cases} \tag{2.1}$$

where the initial condition $\boldsymbol{x}_0$ is either deterministic and known $\hat{\boldsymbol{x}}_0$ or stochastic and drawn from a given density $p_{f,0}(\boldsymbol{x}_0)$. We assume that the aforementioned microscopic evolution law can be integrated in time, albeit with a very small time-step which we denote by δt [†]. For some microscopic systems (e.g. agent-based or chemical-reaction-species models), the evolution laws might be directly prescribed in discretized form.

In order to make the discussion more concrete but also to relate with the systems examined in subsequent sections, we consider in this work microscopic systems which consist of (non)interacting walkers in one-dimension in which case the state vector $\boldsymbol{x}$ represents their coordinates. Hence n_f is also the number of the walkers. It is known [20] that depending on the specifics of the random walks and their interactions, the emergent behavior of the overall FG system can exhibit a wide range of features and time scales (e.g. shock formation, see Section 3). The challenge we undertake is to find an effective, CG model that provides an adequate description of the FG system of walkers.

This is generally approached as a two-step process, where the analyst *separately*:

1. identifies a set of CG variables, say $\boldsymbol{X}\in\mathcal{M}_c\subset\mathbb{R}^{n_c}$, $(n_c\ll n_f)$ through a *fine-to-coarse* map (restriction operator),
2. and constructs a closed CG model i.e. an evolution law for the CG variables e.g. $\dot{\boldsymbol{X}}_t=\boldsymbol{F}(\boldsymbol{X}_t)$.

In order to address the difficulties associated with these tasks, but also to construct a CG model that is capable of predicting, albeit with uncertainty, the whole FG picture, we adopt here a different strategy which is based on the *simultaneous* learning of:

[†]We do not account for the integration errors in this study as this is outside the scope of the paper, see e.g. [62].

- a *probabilistic evolution law* for the CG states $p_c(\boldsymbol{X}_{t+\Delta t}|\boldsymbol{X}_t)$;
- a *probabilistic coarse-to-fine* map $p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t)$.

where $\boldsymbol{X}_t$ denotes the CG state variables at time t and $\Delta t \gg \delta t$ the coarse time-step used for the evolution law of the CG states. We employ here directly the *discretized* evolution law as this will ultimately be used for predictive purposes. We have also prescribed an evolution law that is Markovian in order to simplify the presentation, although this can be relaxed. Before we discuss in detail the aforementioned densities, their parametrization as well as their calibration, we provide some crucial details about the overall formulation.

We note that the CG states are indirectly defined as latent generators that give rise to the FG states through the probabilistic lifting operator implied by p_{cf}. The interpretation of these latent variables can only be done through the prism of this generative mapping. According to this, each sequence of FG configurations $\boldsymbol{x}_t$ at times $t=k\Delta t$, $k=0,1,\cdots,n$ is generated as follows:

- One selects a CG initial state $\boldsymbol{X}_0$, drawn for example from a prior density $p_{c,0}(\boldsymbol{X}_0)$. The initial FG state $\boldsymbol{x}_0$ is drawn from $p_{cf}(\boldsymbol{x}_0|\boldsymbol{X}_0)$.
- For $k=0,1,\cdots,n$:
  - evolve the CG model in time by sampling successive states from $p_c(\boldsymbol{X}_{(k+1)\Delta t}|\boldsymbol{X}_{k\Delta t})$,
  - sample $\boldsymbol{x}_{k\Delta t}$ from $p_{cf}(\boldsymbol{x}_{k\Delta t}|\boldsymbol{X}_{k\Delta t})$.

The aforementioned procedure implies that $\boldsymbol{x}_t$ is effectively drawn from the following (conditional) density (given $\boldsymbol{X}_0$):

$$\begin{aligned}&\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t}|\boldsymbol{X}_0)\\&=\int p_{cf}(\boldsymbol{x}_0|\boldsymbol{X}_0)\left(\prod_{k=0}^{n-1}p_c(\boldsymbol{X}_{(k+1)\Delta t}|\boldsymbol{X}_{k\Delta t})p_{cf}(\boldsymbol{x}_{(k+1)\Delta t}|\boldsymbol{X}_{(k+1)\Delta t})\right)d\boldsymbol{X}_{\Delta t}d\boldsymbol{X}_{2\Delta t}\cdots d\boldsymbol{X}_{n\Delta t}.\end{aligned}\tag{2.2}$$

Naturally if the uncertainty in $\boldsymbol{X}_0$ is incorporated then:

$$\begin{aligned}&\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t})\\&=\int p_{c,0}(\boldsymbol{X}_0)\ p_{cf}(\boldsymbol{x}_0|\boldsymbol{X}_0)\left(\prod_{k=0}^{n-1}p_c(\boldsymbol{X}_{(k+1)\Delta t}|\boldsymbol{X}_{k\Delta t})p_{cf}(\boldsymbol{x}_{(k+1)\Delta t}|\boldsymbol{X}_{(k+1)\Delta t})\right)d\boldsymbol{X}_0\ d\boldsymbol{X}_{\Delta t}\cdots d\boldsymbol{X}_{n\Delta t}.\end{aligned}\tag{2.3}$$

Suppose the aforementioned densities are parametrized by $\boldsymbol{\theta}=(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})$ i.e. $p_c(\boldsymbol{X}_{t+\Delta t}|\boldsymbol{X}_t,\ \boldsymbol{\theta}_c)$ and $p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t,\ \boldsymbol{\theta}_{cf})$ and we attempt to minimize the Kullback-Leibler divergence [63] between the reference density $p_f(\boldsymbol{x}_{(0:n)\Delta t})$ and $\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t}|\boldsymbol{\theta})$. The former is either a Dirac-delta in the case of deterministic initial condition and deterministic FG

dynamics in Eq. (2.1) or a proper density in the case of stochastic initial conditions or stochastic FG evolution law [64]‡. Then:

$$KL(p_f(\boldsymbol{x}_{(0:n)\Delta t})||\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t}|\boldsymbol{\theta}) = -\int p_f(\boldsymbol{x}_{(0:n)\Delta t})\log\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t}|\boldsymbol{\theta})\ d\boldsymbol{x}_{(0:n)\Delta t} + \int p_f(\boldsymbol{x}_{(0:n)\Delta t})\log p_f(\boldsymbol{x}_{(0:n)\Delta t})\ d\boldsymbol{x}_{(0:n)\Delta t}. \tag{2.4}$$

Given that the second term, i.e. the negative entropy of p_f, is independent of $\boldsymbol{\theta}$, minimizing the KL-divergence is equivalent to maximizing [9]:

$$\mathcal{L}(\boldsymbol{\theta}) = \int p_f(\boldsymbol{x}_{(0:n)\Delta t})\log\bar{p}_f(\boldsymbol{x}_{(0:n)\Delta t}|\boldsymbol{\theta})\ d\boldsymbol{x}_{(0:n)\Delta t} \tag{2.5}$$

which can be approximated by sampling N times from $p_f(\boldsymbol{x}_{(0:n)\Delta t})$ i.e. by simulating the FG dynamics up to $n\Delta t$ in order to obtain the FG sequences $\{\boldsymbol{x}^{(i)}_{(0:n)\Delta t}\}_{i=1}^N$. The resulting approximation which we denote by $L_N(\boldsymbol{\theta})$, effectively represents *log-likelihood* of the FG simulated data with respect to the generative model proposed, i.e.,

$$\begin{aligned} L_N(\boldsymbol{\theta}) &= \sum_{i=1}^N \log\bar{p}_f(\boldsymbol{x}^{(i)}_{(0:n)\Delta t}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^N \log\int p_{c,0}(\boldsymbol{X}^{(i)}_0)p_{cf}(\boldsymbol{x}^{(i)}_0|\boldsymbol{X}^{(i)}_0)\left(\prod_{k=0}^{n-1} p_c(\boldsymbol{X}^{(i)}_{(k+1)\Delta t}|\boldsymbol{X}^{(i)}_{k\Delta t})p_{cf}(\boldsymbol{x}^{(i)}_{(k+1)\Delta t}|\boldsymbol{X}^{(i)}_{(k+1)\Delta t})\right) d\boldsymbol{X}^{(i)}_{(0:n)\Delta t}. \end{aligned} \tag{2.6}$$

We note in the expression above that we associate a sequence of latent, CG states $\boldsymbol{X}^{(i)}_{(0:n)\Delta t}$ to each observed FG sequence $\boldsymbol{x}^{(i)}_{(0:n)\Delta t}$. Hence the CG states play the role of pre-images of the FG states. More importantly, the objective in the aforementioned expression accounts for both the CG probabilistic evolution law as well as the reconstruction (lifting) of the FG states from the (latent) CG ones. The $\boldsymbol{\theta}$ that maximize $L_N(\boldsymbol{\theta})$ would obviously correspond to the Maximum Likelihood estimate $\boldsymbol{\theta}_{MLE}$. Furthermore the interpretation of the objective as the log-likelihood makes the passage to Bayesian formulations, straightforward. If for example we define a prior density $p(\boldsymbol{\theta})$, then maximizing:

$$L_N(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{2.7}$$

would be equivalent to obtaining a Maximum A Posteriori (MAP) estimate $\boldsymbol{\theta}_{MAP}$. The natural progression from point estimates, is a fully Bayesian treatment which involves the *posterior density* of $\boldsymbol{\theta}$ that according to Bayes' formula is given by:

$$p(\boldsymbol{\theta}|\ \boldsymbol{x}^{(1:N)}_{(0:n)\Delta t}) = \frac{e^{L_N(\boldsymbol{\theta})}\ p(\boldsymbol{\theta})}{\bar{p}_f(\boldsymbol{x}^{(1:N)}_{(0:n)\Delta t})}. \tag{2.8}$$

‡As we show in the sequel, it is not necessary to know the density p_f (which, e.g. if Fokker-Planck equations need to be solved, poses an even harder problem) but it suffices to be able to sample from it i.e. simulate the FG dynamics.

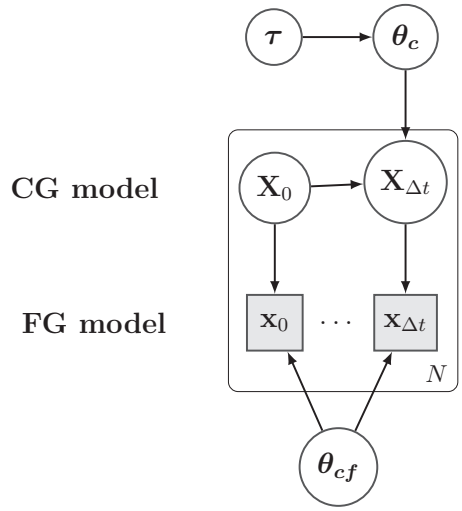


Figure 1: Proposed probabilistic graphical model (for $n=1$). The dotted lines in the evolution of the FG model indicate that in general, the CG time step Δt is a multiple of the FG one δt.

We note that the posterior is well-defined regardless:

- of the type of the FG model (since it relies on simulation data),
- if one ($N=1$) or more ($N>1$) data sequences, with the same or different lengths, are available,
- if these involve one ($n=1$) or more ($n>1$) coarse time-steps Δt.

The aforementioned relationship can be concretely represented in the form of a directed graphical model as depicted in Fig. 1 (for $n=1$).

We discuss in detail a strategy for approximating this posterior in the next subsections. It is more important to emphasize at this stage that given this posterior, we can produce, not just point estimates of FG state futures by simulating the learned CG model, but also quantify the uncertainty in these predictions. This predictive uncertainty reflects the various sources described previously. Let $\boldsymbol{x}_t^{(I)}$ the FG state of data sequence I ($1\leq I\leq N$) at a *future* time $t=(n+m)\Delta t>n\Delta t$ as compared to the data available. The proposed model provides a *predictive posterior* for $\boldsymbol{x}_t^{(I)}$ which can be readily determined by successive applications of the sum and product rules as follows:

$$\begin{aligned} p(\boldsymbol{x}_t^{(I)}|\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)}) &= \int p(\boldsymbol{x}_t^{(I)},\boldsymbol{X}_t^{(I)},\boldsymbol{\theta}|\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\, d\boldsymbol{X}_t^{(I)}\, d\boldsymbol{\theta} \\ &= \int \underbrace{p(\boldsymbol{x}_t^{(I)}|\boldsymbol{X}_t^{(I)},\boldsymbol{\theta},\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})}_{p_{cf}(\boldsymbol{x}_t^{(I)}|\boldsymbol{X}_t^{(I)},\boldsymbol{\theta}_{cf})}\, p(\boldsymbol{X}_t^{(I)}|\boldsymbol{\theta},\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\, \underbrace{p(\boldsymbol{\theta}|\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})}_{\textit{posterior Eq. (2.22)}}\, d\boldsymbol{X}_t^{(I)}\, d\boldsymbol{\theta} \\ &= \int p_{cf}(\boldsymbol{x}_t^{(I)}|\boldsymbol{X}_t^{(I)},\boldsymbol{\theta}_{cf})\, p(\boldsymbol{X}_t^{(I)}|\boldsymbol{\theta},\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\, p(\boldsymbol{\theta}|\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\, d\boldsymbol{X}_t^{(I)}\, d\boldsymbol{\theta}, \end{aligned} \tag{2.9}$$

where $p(\boldsymbol{X}_t^{(I)}|\boldsymbol{\theta},\boldsymbol{X}_{(0:n)\Delta t}^{(1:N)})$ is the (conditional, given $\boldsymbol{\theta}$) predictive posterior of the (latent) CG state of the data sequence I, at the future time t. This can be found using the posterior of $\boldsymbol{X}_{n\Delta t}^{(I)}$ i.e. the posterior of the latent CG state at the time up to which data is available ($n\Delta t$) and by propagating the CG probabilistic evolution law. In particular:

$$\begin{aligned}
&p(\boldsymbol{X}_t^{(I)}|\boldsymbol{\theta},\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\\
&=\int p(\boldsymbol{X}_{(n+m)\Delta t'}^{(I)},\boldsymbol{X}_{(n+m-1)\Delta t'}^{(I)},\cdots,\boldsymbol{X}_{n\Delta t}^{(I)}|\boldsymbol{\theta},\boldsymbol{x}_{(0:n)\Delta t}^{(1:N)})\, d\boldsymbol{X}_{(n+m-1)\Delta t}^{(I)}\cdots d\boldsymbol{X}_{n\Delta t}^{(I)}\\
&=\int \prod_{k=0}^{m-1} p_c(\boldsymbol{X}_{(n+k+1)\Delta t}^{(I)}|\boldsymbol{X}_{(n+k)\Delta t}^{(I)},\boldsymbol{\theta}_c)\; p(\boldsymbol{X}_{n\Delta t}^{(I)}|\boldsymbol{x}_{(0:n)\Delta t}^{(I)},\boldsymbol{\theta})\, d\boldsymbol{X}_{(n+m-1)\Delta t}^{(I)}\cdots d\boldsymbol{X}_{n\Delta t}^{(I)}. \qquad (2.10)
\end{aligned}$$

The combination of Eqs. (2.9) and (2.10) suggests the following procedure in order to sample from the predictive posterior of the future FG state $\boldsymbol{x}_t^{(I)}$:

1. Sample $\boldsymbol{\theta}=(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})$ from the (marginal) posterior in Eq. (2.22).
2. Sample $\boldsymbol{X}_{n\Delta t}^{(I)}$ from the (conditional) posterior $p(\boldsymbol{X}_{n\Delta t}^{(I)}|\boldsymbol{x}_{(0:n)\Delta t}^{(I)},\boldsymbol{\theta})$.
3. Propagate/Sample the CG model until $t=(n+m)\Delta t$ by employing the stochastic evolution law implied by p_c for the $\boldsymbol{\theta}_c$ from step 1.
4. Sample $\boldsymbol{x}_t^{(I)}$ from $p_{cf}(\boldsymbol{x}_t^{(I)}|\boldsymbol{X}_t^{(I)},\boldsymbol{\theta}_{cf})$ for the $\boldsymbol{X}_t^{(I)}$ from step 3 and $\boldsymbol{\theta}_{cf}$ from step 1.

We present in detail how all these steps can be carried out in the next subsections where p_c, p_{cf} and the corresponding posteriors are given specific forms.

2.2 Specification for microscopic random walkers

In this subsection, we present the specific form of the components of the state-space model discussed previously for the particular microscopic models considered. Special emphasis is placed on the parametrization adopted that enables the identification of the CG model's form, the enforcement of macroscopic constraints (in this case, conservation laws), as well as the inference engine employed that makes use of Stochastic Variational Inference methods. We conclude with an algorithm that performs short numerical experiments by appropriately initializing the FG model and ingests the simulation data generated in order to calibrate the CG model. We aim also at providing insight that is generalizable to other problem types.

2.2.1 The coarse-to-fine probabilistic map p_{cf}

The FG models considered in this section consist of identical, random walkers in a bounded subdomain, say $[y_{\min},y_{\max}]\subset\mathbb{R}$ where the FG variables $\boldsymbol{x}_t\in\mathbb{R}^{n_f}$ represent the coordinate of each of the n_f walkers at time t. We will consider different types interactions

between the walkers as well as various initial conditions, but in all cases it is assumed that the FG model can be simulated, albeit with a significant computational cost and with a small time-step δt. Without loss of generality we employ periodic boundary conditions. For such problems, it is clear that the walker-density represents a reasonable CG state variable [20] and normally one would attempt to write an appropriate evolution law for it. As discussed earlier, our starting point is the specification of the CG state variables through a coarse-to-fine map as described by p_{cf}. In particular, if $\rho(y,t)$ denotes the normalized density (i.e. $\int_{y_{\min}}^{y_{\max}} \rho(y,t)\, dy = 1, \forall t$) of the walkers which depends on space $y \in \mathbb{R}$ and time t, let $\boldsymbol{\rho}_t = \{\rho_{t,j}\}_{j=1}^{n_c}$ denote the spatially-discretized vector of the density over $[y_{\min}, y_{\max}]$ with a step size Δy such that $\rho_{t,j}$ expresses the relative number of walkers located in $[y_{\min}+(j-1)\Delta y,\ y_{\min}+j\Delta y]$ at time t, such that at all times:

$$\sum_{j=1}^{n_c} \rho_{t,j} = 1, \quad (\rho_{t,j} \geq 0). \tag{2.11}$$

We note that the equation above enforces a constraint in the evolution of the (discretized) density $\boldsymbol{\rho}_t$ i.e. expresses the *conservation of mass*. Conservation laws constitute the essential building blocks of macroscopic models in continuum thermodynamics [65]. In order to achieve this automatically, i.e. without additional constraints in the CG model, we propose the following softmax transformation:

$$\rho_{t,j} = \frac{e^{X_{t,j}}}{\sum_{l=1}^{n_c} e^{X_{t,l}}} \tag{2.12}$$

where $\boldsymbol{X}_t = \{X_{t,j}\}_{j=1}^{n_c}$ represent the (real-valued) CG state variables employed. Given $\boldsymbol{X}_t$, the coarse-to-fine probabilistic map takes the following form of a multinomial density:

$$p_{cf}(\boldsymbol{x}_t | \boldsymbol{X}_t) = \frac{n_f!}{m_1!\, m_2! \cdots m_{n_c}!} \prod_{j=1}^{n_c} \rho_{t,j}^{m_j}, \tag{2.13}$$

where $m_j = m_j(\boldsymbol{x}_t)$ are the sufficient statistics of $\boldsymbol{x}_t$ that count the number of walkers that at time t are located in the j bin i.e. they are in $[y_{\min}+(j-1)\Delta y,\ y_{\min}+j\Delta y)$. The underlying assumption is that, given the CG state $\boldsymbol{X}_t$ (and consequently $\boldsymbol{\rho}_t$ through Eq. (2.12)), the positions of the walkers, i.e. the FG states, are *conditionally* independent. We emphasize here that this does not imply that the walkers move independently of each other which would be untrue as they can exhibit coherent behavior. This dependence though, is caused indirectly, by the CG states i.e. the same way masses located on separate springs might appear to vibrate independently even when the springs are attached to the same vibrating substrate.

Two remarks are in order at this stage. Firstly, by introducing CG variables that only pertain to the zero-order moment of the particles' distribution (i.e. the density), we have assumed that higher-order moments (e.g. second- (pair) or third-order joint densities or

correlations) are not necessary in predicting the FG states' evolution. The validity of this assumption (in combination with the p_c presented in the next subsection) will be assessed given the FG simulation data. If the CG variables adopted are insufficient to predict the future FG configurations, then this should manifest itself in large predictive uncertainty (e.g. Eq. (2.9)). Methods that focus exclusively on finding closures for the CG variables selected (as discussed in the introduction), even if they are successful in this task, cannot assess *the actual predictive value* of these CG variables and associated models in terms of the original FG system.

For walkers/particles in two or three dimensions, the same CG description could be adopted with the only difference being that two- or three-dimensional bins would need to be employed. Naturally, the density ρ could be represented/discretized differently which could impede though the a-priori enforcement of the mass conservation constraint.

The second observation is that in this particular model, there are no unknown parameters $\boldsymbol{\theta}_{cf}$ that need to be learned from the data. Given the *exchangeability* [66] of the FG state variables $\boldsymbol{x}_t$ arising from the fact that the walkers are identical (i.e. their joint density is invariant to permutations), and according to de Finetti's theorem [67] p_{cf} can be generally expressed as:

$$p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t)=\prod_{j=1}^{n_f} p(x_{t,j}|\boldsymbol{X}_t,\boldsymbol{\theta}_{cf}). \tag{2.14}$$

For non-exchangeable FG variables e.g. as is the case with non-identical walkers/particles, each CG variable could have a different meaning/role in the description. In such cases, any generative, latent variable model [68] would be possible, e.g.

$$p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t)=\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}+\boldsymbol{W}\boldsymbol{X}_t,\boldsymbol{S}), \tag{2.15}$$

where $\boldsymbol{\theta}_{cf}\equiv\{\boldsymbol{\mu},\boldsymbol{W},\boldsymbol{S}\}$ that implies a linear projection on the subspace spanned by the columns of $\boldsymbol{W}$ [69]. Non-linear and non-Gaussian maps could also be employed in order to increase the expressive ability of the model, e.g. (deep) neural nets as in [18,70]. We note though that while such models are rich and flexible due to the large number of parameters, they possess two fundamental drawbacks i.e. they require large amounts of simulation data which might be expensive to obtain for a lot of FG models and lack interpretability of the CG variables $\boldsymbol{X}$.

2.2.2 The CG probabilistic evolution law p_c

The second component of the proposed model involves the specification of a CG evolution law. In traditional formulations this would be in the form of (stochastic) PDE for $\rho(y,t)$ such as:

$$\frac{\partial\rho}{\partial t}=h\Big(t,\rho,\frac{\partial\rho}{\partial y},\frac{\partial^2\rho}{\partial y^2},\cdots\Big). \tag{2.16}$$

This unavoidably raises several questions with regards to the functional form of the right-hand side $h(.)$, the order of the spatial derivatives of ρ that should appear as well the

appropriate values of any parameters in h [20]. The difficulties are only amplified if stochastic terms need to be included. Even if an appropriate CG model was to be identified (as e.g. in simplified cases where closures can be semi-analytically computed directly from the FG model) and was subsequently discretized and integrated forward in time, it would be impossible to retrieve the FG states unless additional assumptions were made. Furthermore, while the discretization errors in the solution of Eq. (2.16) are well-understood, it is unknown how these errors would affect the quality of the predictions for the FG states. Since the discretized model is going to be used for prediction purposes, we also use this in the calibration phase and adopt a probabilistic Markovian model of the hidden CG states $\boldsymbol{X}_t$ of the following form:

$$p_c(\boldsymbol{X}_{(k+1)\Delta t}|\ \boldsymbol{X}_{k\Delta t},\ \boldsymbol{\theta}_c,\ \boldsymbol{v}) = \prod_{j=1}^{n_c} \mathcal{N}(X_{(k+1)\Delta t,\ j}\ ;\ \mu_j(\boldsymbol{X}_{k\Delta t}, \boldsymbol{\theta}_c),\ v_j^{-1}). \tag{2.17}$$

The "deterministic" part in the evolution equations is modeled by the functions μ_j whereas the stochastic fluctuations by the v_j^{-1}'s. Naturally correlated Gaussian or non-Gaussian models can be employed. We propose the following form for the μ_j:

$$\mu_j(\boldsymbol{X}_t, \boldsymbol{\theta}_c) = \boldsymbol{\theta}_c^T \boldsymbol{\phi}^{(j)}(\boldsymbol{X}_t) = \sum_{l=1}^{L} \theta_{c,l}\ \phi_l^{(j)}(\boldsymbol{X}_t). \tag{2.18}$$

Each of the feature functions account for the dependence with the CG states at the previous time step, but one can readily introduce longer memory. Either way, it is obviously impossible to know a priori which $\boldsymbol{\phi}$'s are relevant in the evolution of the CG states, what types of interactions are essential (e.g. first, second-order etc) or how they depend on the coarse-to-fine probabilistic map p_{cf}. This underpins an important *model selection* issue that has been of concern in several coarse-graining studies [71, 72]. One strategy for addressing this, is to initiate the search with a small number of features $\boldsymbol{\phi}$ and progressively add more. These can be selected from a pool of candidates by employing appropriate criteria. In [73] for example, the feature function that causes the largest (expected) decrease (or increase) in the KL-divergence (or the log-likelihood) that we seek to minimize (or maximize), is added at each step. In this work, we adopt a different approach whereby we consider simultaneously a large number of $\boldsymbol{\phi}$. This in turn implies a vector of unknown model parameters $\boldsymbol{\theta}_c$ of very large dimension which not only impedes computations but can potentially lead to multiple local maxima and overfitting particularly in the presence of limited data. More importantly perhaps, it can obstruct the identification of the most salient features of the CG model which provide valuable physical insight [14].

To address this issue, we propose the use of sparsity-inducing priors that are capable of identifying solutions in which only a (small) subset of $\boldsymbol{\theta}_c$ are non-zero and therefore only the corresponding $\boldsymbol{\phi}$'s are active [74]. A lot of the prior models that have been proposed along these lines can be readily cast in the context of hierarchical Bayesian models where hyper-parameters are introduced in the prior. In this work, we adopt the

Automatic Relevance Determination (ARD, [75]) model which consists of the following:

$$p(\boldsymbol{\theta}_c|\boldsymbol{\tau}) = \prod_{l=1}^{L} \mathcal{N}(\theta_{c,l}; 0, \tau_l^{-1}). \tag{2.19}$$

This implies that each $\theta_{c,l}$ is a priori independent, and following a zero-mean, Gaussian with a precision hyper-parameter τ_l. The latter are modeled (independently) with a (conjugate) Gamma i.e.:

$$p(\tau_l) = Gamma(\tau_l; \alpha_0, \beta_0). \tag{2.20}$$

We note that when $\tau_l \to \infty$, then $\theta_{c,l} \to 0$. The resulting prior for $\theta_{c,l}$ arising by marginalizing the hyper-parameter is a heavy-tailed, Student's $t-$distribution [76]. The hyperparameters α_0, β_0 are effectively the only ones that need to be provided by the analyst. We advocate very small values ($\alpha_0 = \beta_0 = 10^{-10}$ was used in the numerical examples) which correspond to a non-informative prior. We also a employed independent Gamma priors for the previsions v_j in Eq. (2.17) as follows:

$$p(v_j) = Gamma(v_j; \gamma_0, \zeta_0), \tag{2.21}$$

with hyperparameters γ_0, ζ_0 which were also set to very small values ($\gamma_0 = \zeta_0 = 10^{-10}$ was used in the numerical examples).

2.3 Generation of training data

Given the aforementioned model, we propose the following procedure for generating training data. It consists of N short, numerical experiments in which the FG model is randomly initialized and propagated for one coarse time-step Δt i.e. for $\frac{\Delta t}{\delta t}$ microscopic time-steps. In particular:

- For $i = 1, \cdots, N$:
  - Sample CG initial state $\boldsymbol{X}_0^{(i)}$ from a density $p_{c,0}(\boldsymbol{X}_0^{(i)})$.
  - Sample FG initial state $\boldsymbol{x}_0^{(i)}$ from $p_{cf}(\boldsymbol{x}_0^{(i)}|\boldsymbol{X}_0^{(i)})$.
  - Solve (discretized) FG model in Eq. (2.1) for $m = \frac{\Delta t}{\delta t}$ microscopic time-steps and record final state $\boldsymbol{x}_{\Delta t}^{(i)}$.

The generated data $\{\boldsymbol{x}_{\Delta t}^{(i)}\}_{i=1}^N$ will be used subsequently to draw inferences on the CG model states and parameters (Section 2.4). We note that longer time sequences could readily be generated (albeit at an increased cost). The number of samples N is also something that can be selected adaptively since as we show in the next sections, inferences and predictions can be updated as soon as more data become available. Furthermore, the predictive estimates produced reflect the informational content of the data. Finally, we note with regards to the density $p_{c,0}(\boldsymbol{X}_0^{(i)})$ from which initial CG states are drawn,

that this can be selected quite flexibly. One can also envisage considering parametrized versions of $p_{c,0}$ where the optimal parameter values can be determined by maximizing information gain criteria in the context of optimal experimental design.

2.4 Inference

In the section we discuss the Bayesian calibration of the proposed model given the synthetic training $\boldsymbol{x}_{\Delta t}^{(1:N)}$ data generated previously. Apart from the parameters $\boldsymbol{\theta}_c$ (and the corresponding hyperparameters $\boldsymbol{\tau}$) and $v=\{v_j\}$ that control the CG model form, the latent CG states $\boldsymbol{X}_{\Delta t}^{(1:N)}$ must be also inferred. As mentioned in Section 2.3, the initial CG states $\boldsymbol{X}_0^{(1:N)}$ are known and are omitted in the ensuing conditional densities for simplicity. The joint posterior is given by:

$$\begin{aligned} p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v}|\boldsymbol{x}_{\Delta t}^{(1:N)}) &= \frac{p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\ p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}{p(\boldsymbol{x}_{\Delta t}^{(1:N)})} \\ &= \frac{p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)})\ p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}{p(\boldsymbol{x}_{\Delta t}^{(1:N)})}. \end{aligned} \tag{2.22}$$

With regards to the terms, we note that the likelihood $p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)})$ can be expressed as:

$$\begin{aligned} p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)}) &= \prod_{i=1}^{N} p(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)}) \\ &= \prod_{i=1}^{N} p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)}), \end{aligned} \tag{2.23}$$

where each of the terms in the product are given by Eq. (2.13). With regards to the prior $p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})$ it decomposes as follows:

$$p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v}) = p(\boldsymbol{X}_{\Delta t}^{(1:N)}|\ \boldsymbol{\theta}_c,\boldsymbol{v})\ p(\boldsymbol{\theta}_c|\ \boldsymbol{\tau})\ p(\boldsymbol{\tau})\ p(\boldsymbol{v}), \tag{2.24}$$

where the last three terms are specified in Eqs. (2.19), (2.20) and (2.21) respectively. Furthermore, the first term depends on the CG evolution law in Eq. (2.17) as:

$$\begin{aligned} p(\boldsymbol{X}_{\Delta t}^{(1:N)}|\ \boldsymbol{\theta}_c,\boldsymbol{v}) &= \prod_{i=1}^{N} p(\boldsymbol{X}_{\Delta t}^{(i)}|\ \boldsymbol{\theta}_c,\boldsymbol{v}) \\ &= \prod_{i=1}^{N} p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\ \boldsymbol{\theta}_c,\boldsymbol{v}). \end{aligned} \tag{2.25}$$

As exact inference is impossible and Monte Carlo-based schemes cumbersome, we attempt to find an approximation $q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})$ to the posterior $p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau}|\ \boldsymbol{x}_{\Delta t}^{(1:N)})$.

To that end, we adopt a Variational Inference scheme [77, 78] which attempts to find the best approximation within an appropriately selected family of densities such that it minimizes the Kullback-Leibler divergence with the posterior i.e.

$$\begin{aligned}
&KL(q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v}) \;||\; p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau}|\,\boldsymbol{x}_{\Delta t}^{(1:N)}))\\
=&-\int q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\log\frac{p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v}|\,\boldsymbol{x}_{\Delta t}^{(1:N)})}{q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}\,d\boldsymbol{X}_{\Delta t}^{(1:N)}\,d\boldsymbol{\theta}_c\,d\boldsymbol{\tau}\,d\boldsymbol{v}\\
=&-\int q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\log p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)})\,d\boldsymbol{X}_{\Delta t}^{(1:N)}\,d\boldsymbol{\theta}_c\,d\boldsymbol{\tau}\,d\boldsymbol{v}\\
&-\int q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\log\frac{p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}{q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}\,d\boldsymbol{X}_{\Delta t}^{(1:N)}\,d\boldsymbol{\theta}_c\,d\boldsymbol{\tau}\,d\boldsymbol{v}+\log p(\boldsymbol{x}_{\Delta t}^{(1:N)}).
\end{aligned}\tag{2.26}$$

Given that the KL-divergence is always non-negative we obtain that:

$$\begin{aligned}
\log p(\boldsymbol{x}_{\Delta t}^{(1:N)})\geq&\int q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\log p(\boldsymbol{x}_{\Delta t}^{(1:N)}|\boldsymbol{X}_{\Delta t}^{(1:N)})\,d\boldsymbol{X}_{\Delta t}^{(1:N)}\,d\boldsymbol{\theta}_c\,d\boldsymbol{\tau}\,d\boldsymbol{v}\\
&+\int q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\log\frac{p(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}{q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})}\,d\boldsymbol{X}_{\Delta t}^{(1:N)}\,d\boldsymbol{\theta}_c\,d\boldsymbol{\tau}\,d\boldsymbol{v}\\
=&\mathcal{F}\Big(q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})\Big).
\end{aligned}\tag{2.27}$$

Hence, finding the q that maximizes the log-evidence lower-bound $\mathcal{F}$ (frequently referred to as ELBO, for Evidence Lower BOund) is equivalent to minimizing the KL-divergence, for which the optimal q being the exact posterior. We adopt a mean-field factorization [79] of the form:

$$q(\boldsymbol{X}_{\Delta t}^{(1:N)},\boldsymbol{\theta}_c,\boldsymbol{\tau},\boldsymbol{v})=\prod_{i=1}^{N}q(\boldsymbol{X}_{\Delta t}^{(i)})\;\;q(\boldsymbol{\theta}_c)\;q(\boldsymbol{\tau})\;q(\boldsymbol{v}).\tag{2.28}$$

This is turn gives rises to an iterative scheme where each of the densities above are updated, until convergence, while the others are kept fixed. It can be readily shown [80] that in this case (based on Eqs. (2.23), (2.24), (2.25)), the optimal q^{opt} i.e. the ones that maximize the ELBO $\mathcal{F}$ are given by (up to a constant):

$$\begin{aligned}
\log q^{opt}(\boldsymbol{\theta}_c)=&<\log p(\boldsymbol{X}_{\Delta t}^{(1:N)}|\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})\;q(\boldsymbol{v})}+<\log p(\boldsymbol{\theta}_c|\boldsymbol{\tau}>_{q(\boldsymbol{\tau})}\\
=&\sum_{i=1}^{N}<\log p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})\;q(\boldsymbol{v})}+<\log p(\boldsymbol{\theta}_c|\boldsymbol{\tau})>_{q(\boldsymbol{\tau})},
\end{aligned}\tag{2.29}$$

$$\log q^{opt}(\boldsymbol{\tau})=<\log p(\boldsymbol{\theta}_c|\boldsymbol{\tau}>_{q(\boldsymbol{\theta})}+\log p(\boldsymbol{\tau}),\tag{2.30}$$

$$\begin{aligned}
\log q^{opt}(\boldsymbol{v})=&<\log p(\boldsymbol{X}_{\Delta t}^{(1:N)}|\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})\;q(\boldsymbol{\theta}_c)}+\log p(\boldsymbol{v})\\
=&\sum_{i=1}^{N}<\log p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})\;q(\boldsymbol{\theta}_c)}+\log p(\boldsymbol{v}),
\end{aligned}\tag{2.31}$$

and

$$\begin{aligned}\log q^{opt}(\boldsymbol{X}_{\Delta t}^{(i)}) &= \log p(\boldsymbol{x}_{\Delta t}^{(i}|\boldsymbol{X}_{\Delta t}^{(i)}) + <\log p(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{\theta}_c)>_{q(\boldsymbol{\theta}_c)} \\ &= \log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)}) + <\log p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\boldsymbol{\theta}_c)>_{q(\boldsymbol{\theta}_c)},\end{aligned} \tag{2.32}$$

where the notation $<.>_q$ implies an expectation with respect to the q density in the subscript. The equations above are obviously coupled, but the first two attain a closed-form solution. In particular, one finds that $q^{opt}(\boldsymbol{\theta}_c)=\mathcal{N}(\boldsymbol{\theta}_c;\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{S}_{\boldsymbol{\theta}})$ where:

$$\begin{aligned}\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} &= \sum_{j=1}^{n_c} <v_j>_{q(\boldsymbol{v})} \sum_{i=1}^{N} \boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0^{(i)})(\boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0^{(i)}))^T + <\boldsymbol{T}>_{q(\boldsymbol{\tau})}, \\ \boldsymbol{S}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} &= \sum_{j=1}^{n_c} <v_j>_{q(\boldsymbol{v})} \sum_{i=1}^{N} \boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0^{(i)}) <X_{\Delta t,\,j}>_{q(\boldsymbol{X}_{\Delta t}^{(i)})},\end{aligned} \tag{2.33}$$

and $\boldsymbol{T}=diag(\tau_l)$. Furthermore for the $\boldsymbol{\tau}$'s, one finds that $q^{opt}(\boldsymbol{\tau})=\prod_l Gamma(\tau_l;\,\alpha_l,\beta_l)$ where:

$$\begin{aligned}\alpha_l &= \alpha_0 + \frac{1}{2}, \\ \beta_l &= \beta_0 + \frac{1}{2} <\theta_{c,j}^2>_{q(\boldsymbol{\theta}_c)} = \beta_0 + \frac{1}{2}(\mu_{\boldsymbol{\theta},l}^2 + \boldsymbol{S}_{\boldsymbol{\theta},ll}).\end{aligned} \tag{2.34}$$

Note that in this case $<\tau_l>_{q(\boldsymbol{\tau})}=\frac{\alpha_l}{\beta_l}$. Similarly, one notes that $q^{opt}(\boldsymbol{v})=\prod_j q_j(v_j)=\prod_j Gamma(v_j;\,\gamma_j,\zeta_j)$ where:

$$\begin{aligned}\gamma_j &= \gamma_0 + \frac{N}{2}, \\ \zeta_j &= \zeta_0 + \frac{1}{2} <\sum_{i=1}^{N} (X_{\Delta t,\,j} - \boldsymbol{\theta}_c^T \boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0^{(i)}))^2>_{q(\boldsymbol{X}_{\Delta t}^{(i)})\; q(\boldsymbol{\theta}_c)}.\end{aligned} \tag{2.35}$$

Note that in this case $<v_j>_{q(\boldsymbol{v})}=\frac{\gamma_j}{\zeta_j}$.

The most challenging component of the update equations pertains to $q^{opt}(\boldsymbol{X}_{\Delta t}^{(i)})$ due to the form of the likelihood (Eq. (2.13)) as well as the softmax transformation in Eq. (2.12) that enforces the conservation of mass. Consequently, it is impossible to derive in closed form q^{opt}. To address this we make use of Stochastic Variational Inference methods [57, 58]. The underlying idea is to approximate the expectations involved with Monte Carlo. In particular, if one denotes with $\mathcal{F}_i$ the terms in $\mathcal{F}$ pertaining to $q(\boldsymbol{X}_{\Delta t}^{(i)})$ from Eq. (2.27), then:

$$\begin{aligned}\mathcal{F}_i = &<\log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})} + <\log p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})q(\boldsymbol{\theta}_c)q(\boldsymbol{v})} \\ &- <\log q(\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})}.\end{aligned} \tag{2.36}$$

In our formulation, we assume a multivariate Gaussian (with a full-rank covariance matrix) for each $q(\boldsymbol{X}_{\Delta t}^{(i)})$, i.e.

$$q(\boldsymbol{X}_{\Delta t}^{(i)}) = \mathcal{N}(\boldsymbol{X}_{\Delta t}^{(i)}; \boldsymbol{\mu}_i, \boldsymbol{S}_i). \tag{2.37}$$

Subsequently, we approximate the derivatives of $\mathcal{F}_i$ with respect to $\boldsymbol{\mu}_i$ and $\boldsymbol{S}_i$ using Monte Carlo, and update these parameters using stochastic gradient ascent [81]. In particular we made use of the *ADAM* algorithm [82], which is one of the most robust, first-order stochastic optimization techniques. A critical role in the lower-variance estimation of the derivatives, was played by the reparametrization trick [83]. Details of this step are contained in Appendix B.

The steps involved in the proposed Variational Inference scheme are summarized in Algorithm 1. We also note that, as one would guess due to the resemblance with Expecation-Maximization schemes [84], various relaxations are possible [85]. For example it is not necessary that the optimal $q(\boldsymbol{X}_{\Delta t}^{(i)})$ (given $q(\boldsymbol{\theta}_c)$ and $q(\boldsymbol{\tau})$) is found at each stage and it suffices to do one or more steps along the (stochastic) gradients of $\mathcal{F}_i$. Furthermore, it is not necessary that one updates the $q(\boldsymbol{X}_{\Delta t}^{(i)})$ for all data points i. It suffices to do so for one or a subset of the data. Finally, it is not necessary that all the data are introduced simultaneously, but it suffices to incorporate them in batches, updating $q(\boldsymbol{\theta}_c)$ and $q(\boldsymbol{\tau})$ each time. Such a scheme provides also a natural tempering effect for the posterior [86]. In the case of very large data sets, which as shown in Section 3 is not necessary here, further stochastic approximations are possible [87].

As a final remark with regards to the Variational Inference perspective adopted, we note that (given the form of the approximate posteriors q) one can approximate (up to Monte Carlo noise) the ELBO $\mathcal{F}$ in Eq. (2.27). This is useful in terms of monitoring convergence during the iterative updates in Algorithm 1 but more importantly perhaps, it allows to approximate the log model evidence $\log p(\boldsymbol{x}_{\Delta t}^{(1:N)})$. The latter is an essential quantitative indicator for scoring competing models (e.g. with different parametrizations) and could serve as the basic objective for model enrichment. We intend to investigate this in future work. We include here the final expression of the ELBO $\mathcal{F}$ based on Eq. (2.27) (up to a constant) when using the optimal $q's$ presented earlier. A detailed derivation is contained in Appendix A.

$$\begin{aligned}\mathcal{F} = &< \sum_{i=1}^{N} \log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)} | \boldsymbol{X}_{\Delta t}^{(i)}) >_{q(\boldsymbol{X}_{\Delta t}^{(i)})} + \frac{1}{2} \sum_{i=1}^{N} \log |\boldsymbol{S}_i| + \frac{1}{2} \log |\boldsymbol{S}_{\boldsymbol{\theta}}| \\ &- \sum_k \alpha_k \log \beta_k - \sum_j \gamma_j \log \zeta_j. \end{aligned} \tag{2.38}$$

2.5 Probabilistic prediction

We devote a few lines to specialize the general formulas given in Section 2.1 in terms of the predictive ability of the proposed model. Consider first one of the data points $x_{\Delta t}^{(I)}$ generated in Section 2.3, whose future evolution we wish to predict (obviously without

Algorithm 1: Proposed EM-algorithm for hidden state and model parameter inference using Stochastic Variational Bayes [88], Automatic Relevance Determination, and stochastic optimization (ADAM [82]).

Data: $\{\boldsymbol{x}_{\Delta t}^{(i)}\}_{i=1}^{N}$
Output: $\{\boldsymbol{\mu}_i,\boldsymbol{S}_i\}_{i=1}^{N}$, $\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{S}_{\boldsymbol{\theta}}$, $\{\alpha_l,\beta_l\}_{l=1}^{L}$, $\{\gamma_j,\zeta_j\}_{j=1}^{n_c}$
Initialize $\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{S}_{\boldsymbol{\theta}}$, $\{\alpha_l,\beta_l\}_{l=1}^{L}$,$\{\gamma_j,\zeta_j\}_{j=1}^{n_c}$;
while $\{\boldsymbol{\mu}_i,\boldsymbol{S}_i\}_{i=1}^{N}$, $\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{S}_{\boldsymbol{\theta}}$, $\{\alpha_l,\beta_l\}_{l=1}^{L}$, $\{\gamma_j,\zeta_j\}_{j=1}^{n_c}$ *not converged* **do**
  for $i\leftarrow 1$ **to** N **do**
    Update $\boldsymbol{\mu}_i,\boldsymbol{S}_i$ by maximizing $\mathcal{F}_i$ (Eq. (2.36)) using ADAM (Appendix B).
  end
  Update $\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{S}_{\boldsymbol{\theta}}$ based on Eq. (2.33);
  Update $\{\alpha_l,\beta_l\}_{l=1}^{L}$ based on Eq. (2.34);
  Update $\{\gamma_j,\zeta_j\}_{j=1}^{n_c}$ based on Eq. (2.35);
  Estimate $\mathcal{F}$ based on Eq. (2.38) (Appendix A) ;
end

employing the FG model). Let $q(\boldsymbol{X}_{\Delta t}^{(I)})$ be the (approximate) posterior of the corresponding hidden CG state and $q(\boldsymbol{\theta}_c)$ the (approximate) posterior for $\boldsymbol{\theta}_c$ (that depends on all training data $\boldsymbol{x}_{\Delta t}^{(1:N)}$ as seen in Eq. (2.29)). Then our model provides a *predictive* posterior density for $x_t^{(I)}$ at any future time $t=n\Delta t$ ($n>1$). In particular:

$$\begin{aligned} &p(\boldsymbol{x}_{n\Delta t}^{(I)}|\boldsymbol{x}_{\Delta t}^{(1:N)}) \\ &=\int p_{cf}(\boldsymbol{x}_{n\Delta t}^{(I)}|\boldsymbol{X}_{n\Delta t}^{(I)})\left(\prod_{k=1}^{n-1}p_c(\boldsymbol{X}_{(k+1)\Delta t}^{(I)}|\boldsymbol{X}_{k\Delta t}^{(I)},\boldsymbol{\theta}_c)\right) q(\boldsymbol{X}_{\Delta t}^{(I)})\, q(\boldsymbol{\theta}_c)\, d\boldsymbol{X}_{n\Delta t}^{(I)}\cdots d\boldsymbol{X}_{\Delta t}^{(I)}\, d\boldsymbol{\theta}_c, \end{aligned} \tag{2.39}$$

where p_{cf} and p_c are provided by Eqs. (2.13) and (2.17) respectively.

The second case we finally consider is that if *new* (not included in the training data) FG state $\bar{\boldsymbol{x}}_0$ whose future evolution we would like to predict using the CG model trained on the data $\boldsymbol{x}_{\Delta t}^{(1:N)}$ discussed previously. We do not discuss a longer FG sequence (e.g. $\bar{\boldsymbol{x}}_0$ and $\bar{\boldsymbol{x}}_{\Delta t}$) as this would be incorporated in the training data as above[§]. The first step in this case would be to find the posterior $p(\bar{\boldsymbol{X}}_0|\bar{\boldsymbol{x}}_0)$ which is proportional to:

$$p(\bar{\boldsymbol{X}}_0|\bar{\boldsymbol{x}}_0)\propto p_{cf}(\bar{\boldsymbol{x}}_0|\bar{\boldsymbol{X}}_0)\; p_{c,0}(\bar{\boldsymbol{X}}_0), \tag{2.40}$$

where $p_{c,0}$ is a prior that can be taken to be e.g. a vague Gaussian (in the case of $n_f \gg n_c$ its effect is minimal). Given this (it can also be approximated using SVI as before), the

[§]The only fundamental difference in this case is that the CG initial state $\bar{\boldsymbol{X}}_0$ is unknown and would need to be inferred in addition to $\bar{\boldsymbol{X}}_{\Delta t}$. This joint inference with the goal of finding a joint approximate posterior $q(\bar{\boldsymbol{X}}_0,\bar{\boldsymbol{X}}_{\Delta t})$, can be carried out using the aforementioned framework e.g. by assuming a joint Gaussian and optimizing the joint mean and covariance.

predictive posterior for the evolution at any time $t=n\Delta t$ is given by:

$$p(\bar{\boldsymbol{x}}_{n\Delta t}|\bar{\boldsymbol{x}}_0,\, \boldsymbol{x}_{\Delta t}^{(1:N)}) \\ = \int p_{cf}(\bar{\boldsymbol{x}}_{n\Delta t}|\bar{\boldsymbol{X}}_{n\Delta t}) \left(\prod_{k=0}^{n-1} p_c(\bar{\boldsymbol{X}}_{(k+1)\Delta t}|\bar{\boldsymbol{X}}_{k\Delta t},\boldsymbol{\theta}_c)\right)\, p(\bar{\boldsymbol{X}}_0|\bar{\boldsymbol{x}}_0)\, q(\boldsymbol{\theta}_c)\, d\bar{\boldsymbol{X}}_{n\Delta t}\cdots d\bar{\boldsymbol{X}}_0\, d\boldsymbol{\theta}_c. \quad (2.41)$$

3 Numerical illustrations

In this section, we present three numerical illustrations demonstrating various aspects and capabilities of the proposed framework. Before embarking in the presentation of the results, we provide details of the feature functions $\boldsymbol{\phi}$ (Eq. (2.18)) employed which are common in all three cases. These provide the vocabulary with which the CG evolution is expressed. In particular, we consider feature functions implying first-order interactions:

$$\phi_m^{(j)}(\boldsymbol{X}) = X_{j+m}, \quad m=-M,\cdots,0,\cdots M, \quad (3.1)$$

as well as second order, i.e.

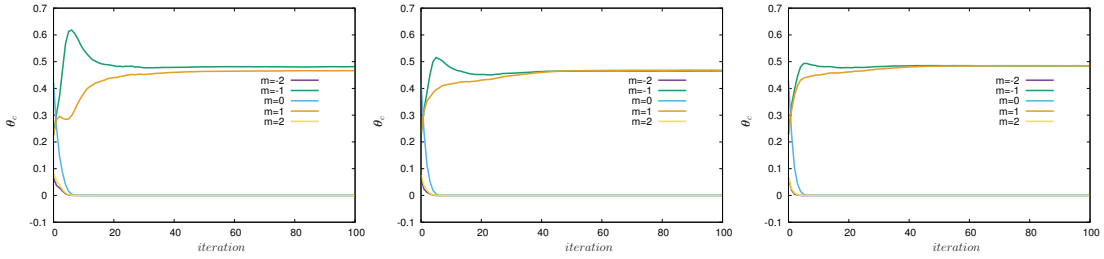
$$\phi_{m_1,m_2}^{(j)}(\boldsymbol{X}) = X_{j+m_1}X_{j+m_2}, \quad m_1,m_2=-M,\cdots,0,\cdots M. \quad (3.2)$$

Higher-order feature functions of polynomial form or functions of different form (e.g. radial basis kernels) can also be employed but in the examples considered were never activated and are therefore omitted from this discussion. Given the discretized nature of the CG evolution law, we note that the interaction range implied by M effectively defines the highest order of spatial derivatives which can be represented with finite differences (see e.g. Eq. (2.16)). The second-order feature functions correspond to nonlinear terms that might be present in the right-hand of the CG evolution law (Eq. (2.16)). We also employed the same variances $v_j=v, \forall j=1,\cdots,n_c$ for all CG state variables (Eq. (2.17)) and report this as v in the following.

We note finally, that as we mentioned earlier for p_{cf}, over-parametrized models such as (deep) neural nets or non-parametric models such as Gaussian processes could, in principle be employed here to represent the right-hand side of the CG evolution law. These however would either require large amounts of FG simulation data for training and/or lack the interpretability that the aforementioned feature functions afford.

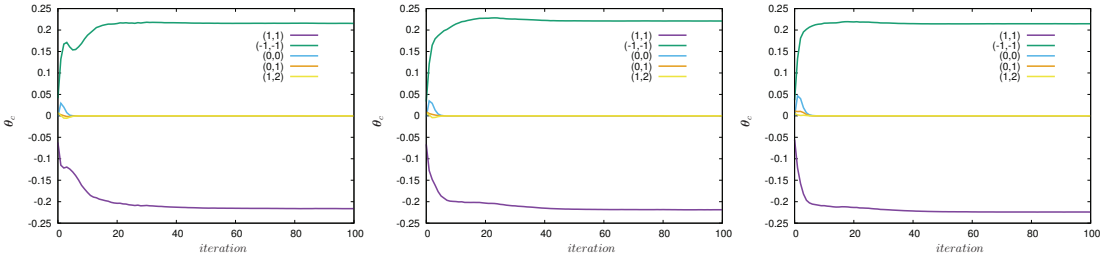
3.1 Synthetic example

The primary goal of this example is to assess the accuracy of the inference engine employed, the effect of the number of training data, the capability of the ARD prior in identifying sparse solutions and to provide insight to the proposed algorithm. To that end, we assumed values of the parameters of the CG evolution law (i.e. $\boldsymbol{\theta}_c,\boldsymbol{v}$ in Eq. (2.17)) and generated FG data as described in Section 2.3. The data produced (for $dim(\boldsymbol{X})=n_c=24$



(a) $N=64$ (b) $N=128$ (c) $N=256$

Figure 2: Posterior mean of $\boldsymbol{\theta}_c$ associated with first-order feature functions (Eq. (3.1)) per iteration of Algorithm 1.



(a) $N=64$ (b) $N=128$ (c) $N=256$

Figure 3: Posterior mean of $\boldsymbol{\theta}_c$ associated with (some) second-order feature functions (Eq. (3.2)) per iteration of Algorithm 1.

and number of walkers $n_f=4800$) were subsequently used in order to assess whether the ground truth could be recovered.

In particular, we considered $M=2$ (Eqs. (3.1), (3.2)) and set all $\boldsymbol{\theta}_c$ coefficients to zero, except for:

- $m=-1$ and $m=+1$ in Eq. (3.1) for which the value 0.5 was assigned;
- $(m_1=1,\ m_2=1)$ in Eq. (3.2) for which the value -0.23 was assigned and for $(m_1=-1,\ m_2=-1)$ the value 0.21.

Hence, 26 out of the 30 features available were turned off. We also assumed a deterministic evolution law i.e. set $v^{-1}=v_j^{-1}=0$.

Figs. 2 and 3 depict the evolution of the posterior means of $\boldsymbol{\theta}_c$ associated with first-order and (some) second-order feature functions for three data sizes, namely $N=$ 64,128,256. We observe that in all cases the ground-truth values are fairly accurately recovered and the level of accuracy improves with increasing data, as expected. More importantly perhaps, the model correctly identifies the active features and the θ_c associated with the inactive ones are set to zero (this also holds for the second-order features not depicted in Fig. 3 due to space limitations). Fig. 4 depicts the evolution of the reciprocal of the posterior mean $<v>$ of the CG model's precision (Eq. (2.17)) which converges

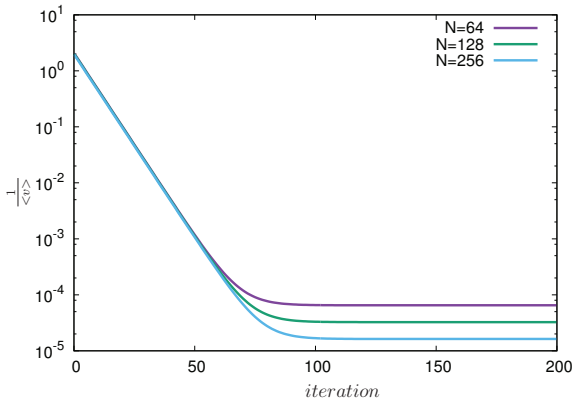
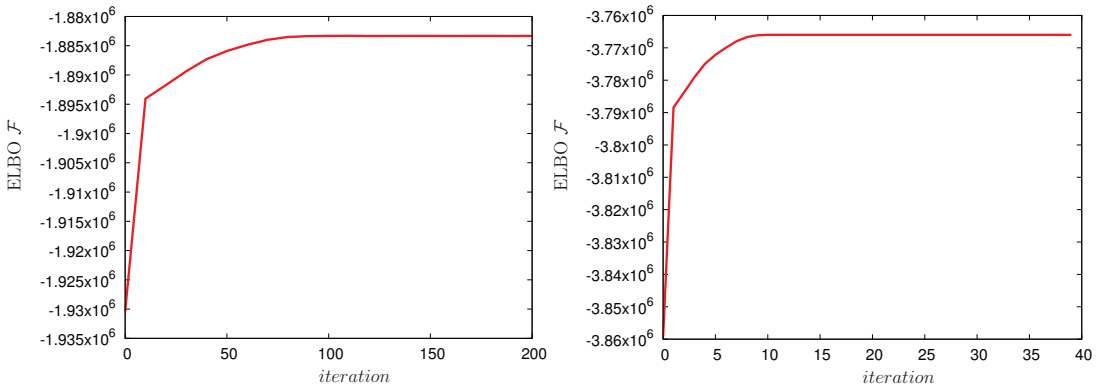


Figure 4: Reciprocal of the posterior mean of v per iteration of Algorithm 1.



(a) $N=128$ (b) $N=256$

Figure 5: Evolution of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38) for $N=128$ (left) and $N=256$ (right).

to very small values approaching the reference value 0. Fig. 5 depicts the ascending evolution (as expected) of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38), for $N=128$ and 256 (qualitatively this is the same as for other N although the range of values can differ drastically). Fig. 6 compares the true CG state vector $\boldsymbol{X}_{\Delta t}$ with the results of the Variational posterior approximation which is trained purely on FG data. One can clearly observe very good agreement and tight uncertainty bounds as reflected in the small posterior variances.

The remaining figures pertain to predictive estimates. In particular, in Fig. 7, we compare the evolution of the walker density at various future times and for two different initial profiles with the predictive posterior estimates (Section 2.5). One observes that the 95% credible interval almost always envelops the reference, even when the reference profile corresponds to all the walkers being concentrated in a single bin. In these illustrations, we employed the same number of bins (24) as those used in the CG description

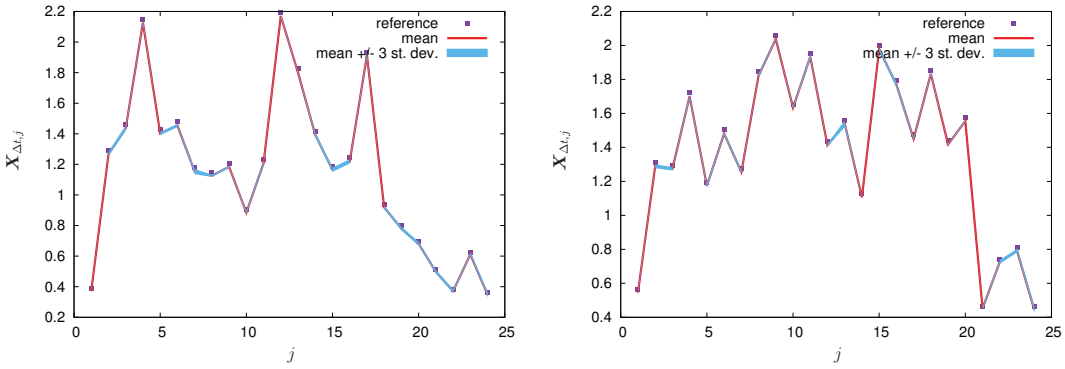


Figure 6: Comparison of reference $X_{\Delta t,j}$ with posterior mean and standard deviation as inferred using FG data using Algorithm 1, for two different cases.

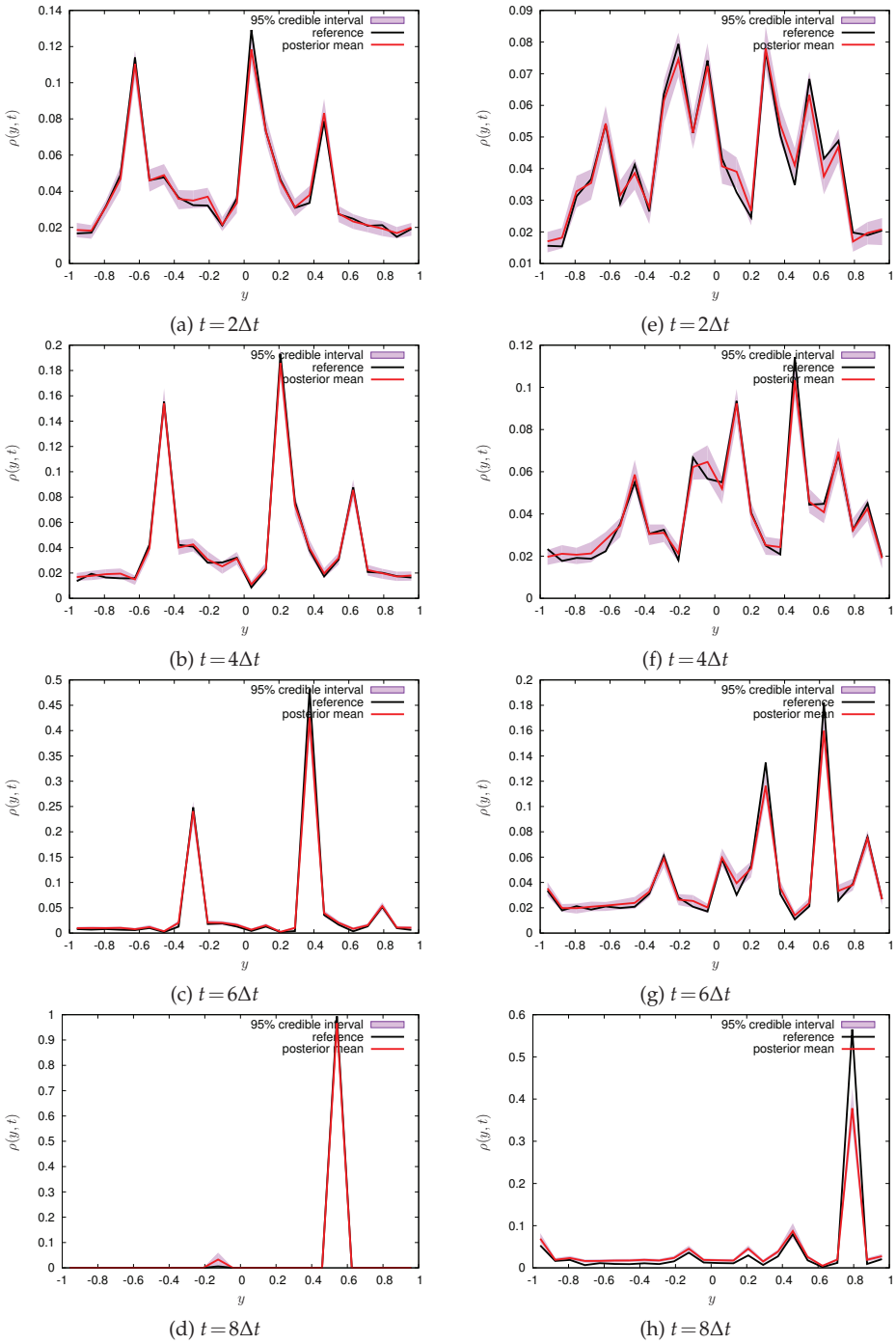
even though the predictive estimates are based on a full reconstruction of the FG states $\boldsymbol{x}_t$ as described in Section 2.5 and not the CG states $\boldsymbol{X}_t$. To further demonstrate this aspect of the model, in Fig. 8, we depict the same predictive estimates of the walker's density as reconstructed from the FG states $\boldsymbol{x}_t$ and using a binning that is 4 times finer than the CG states (i.e. 96 bins). Again, highly accurate estimates are observed with one important difference. The credible intervals are wider i.e. reflect increased predictive uncertainty. This can be explained by the fact that when finer level details are sought, it is unavoidable that greater uncertainty is present.

Finally, in Fig. 9 we depict predictive estimates pertaining to second-order statistics. In particular, we focus on the probability of finding two walkers, simultaneously, at two bins (k_1,k_2) (we call this 2-bin probability) and compare reference values (i.e. those obtained by simulating the FG model) with the probabilistic predictive estimates of the proposed model. We do so for some sample values of (k_1,k_2), for different time instants in the future, and for two different numbers of bins i.e. 24 (left column) and 96 (right column). Despite the fact that the CG model pertains to the walker density, accurate predictive estimates of the second-order FG statistics can still be obtained. As before, the predictive uncertainty is increased when a finer binning is employed.

3.2 Advection-diffusion example

The next system considered consists at the fine-scale of n_f walkers which perform random jumps by δy either to left or right with probabilities $k_l \cdot \delta t$ and $k_r \cdot \delta t$ (no interaction). Therefore the probability of staying at the same position is $1-(k_l+k_r)\delta t$. It is well-known ([89]) that, in the limit $(n_f \to \infty, \delta x, \delta t \to 0)$, this yields an advection-diffusion process with a diffusion constant $D=(k_l+k_r)\frac{\delta y^2}{2\delta t}$ and an advection velocity of $v_a = \frac{k_r-k_l}{\delta t}$ ¶. Clearly when $k_r = k_l$ a pure diffusion is obtained. In the following example, we employed $n_f = 2400$ walkers, $\delta y = 3.875\times 10^{-3}$, $\delta t = 2.5\times 10^{-3}$. The latter two numbers should be compared

¶In this example, the following values were used $k_l = 0.195$, $k_r = 0.205$



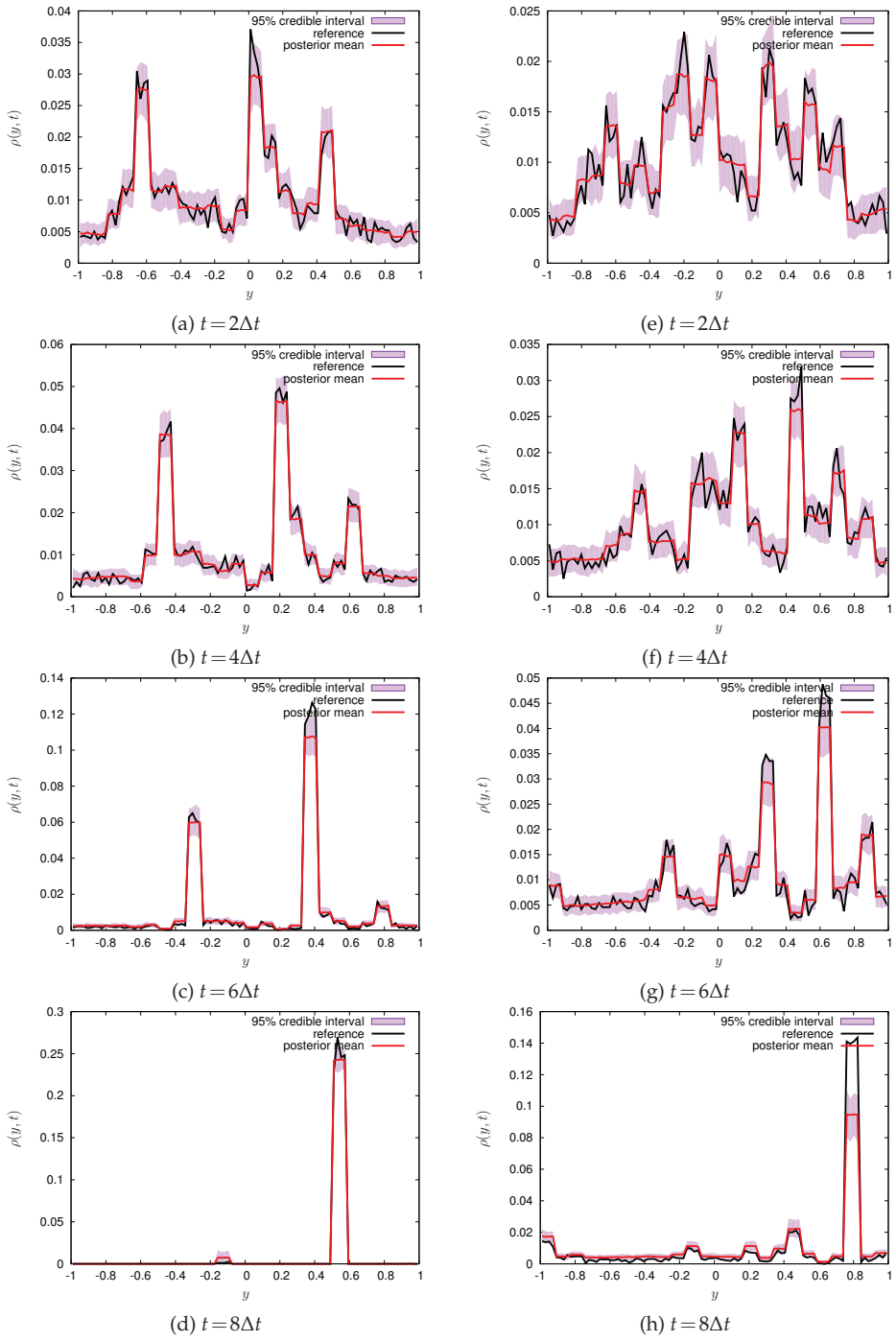
(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=4\Delta t$ (f) $t=4\Delta t$

(c) $t=6\Delta t$ (g) $t=6\Delta t$

(d) $t=8\Delta t$ (h) $t=8\Delta t$

Figure 7: Predictive estimates of walker density for 24 bins (same as in CG evolution), for various times in the future i.e. $t=2\Delta t,4\Delta t,6\Delta t,8\Delta t$ for two different initial conditions (columns). The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=256$)



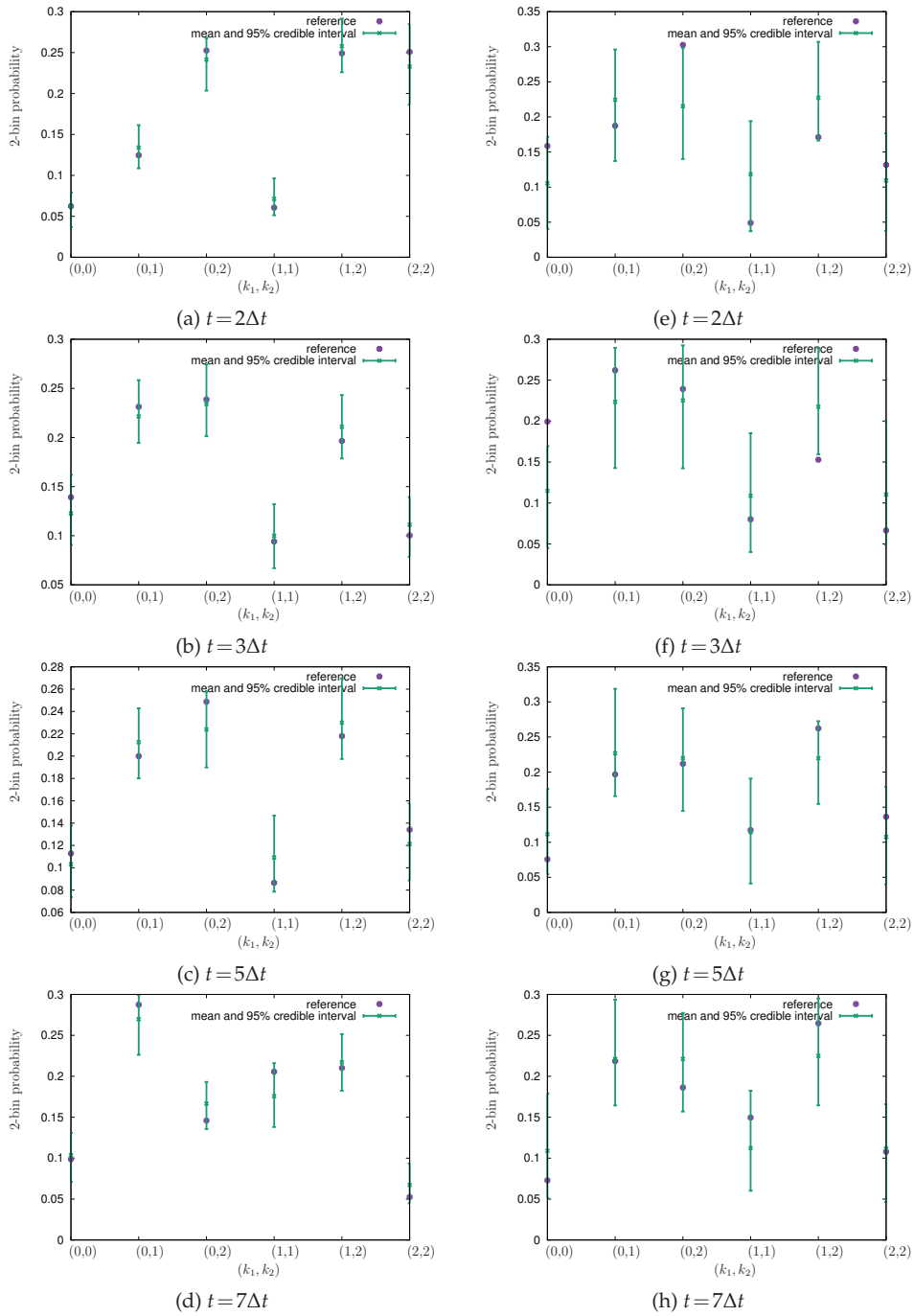
(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=4\Delta t$ (f) $t=4\Delta t$

(c) $t=6\Delta t$ (g) $t=6\Delta t$

(d) $t=8\Delta t$ (h) $t=8\Delta t$

Figure 8: Predictive estimates of walker density *for* 96 *bins* (4 times greater than the 24 used in CG evolution), for various times in the future i.e. $t=2\Delta t,4\Delta t,6\Delta t,8\Delta t$ for two different initial conditions (columns). The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=256$)



(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=3\Delta t$ (f) $t=3\Delta t$

(c) $t=5\Delta t$ (g) $t=5\Delta t$

(d) $t=7\Delta t$ (h) $t=7\Delta t$

Figure 9: Predictive estimates of 2-bin probability for various times in the future i.e. $t=2\Delta t,3\Delta t,5\Delta t,7\Delta t$. The left column corresponds to the original binning, whereas the right column to 4 times finer binning. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=256$)

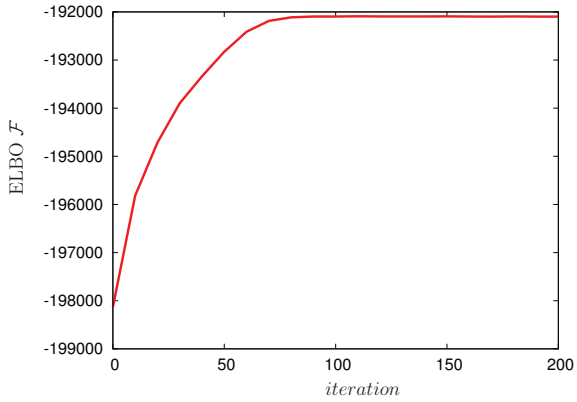
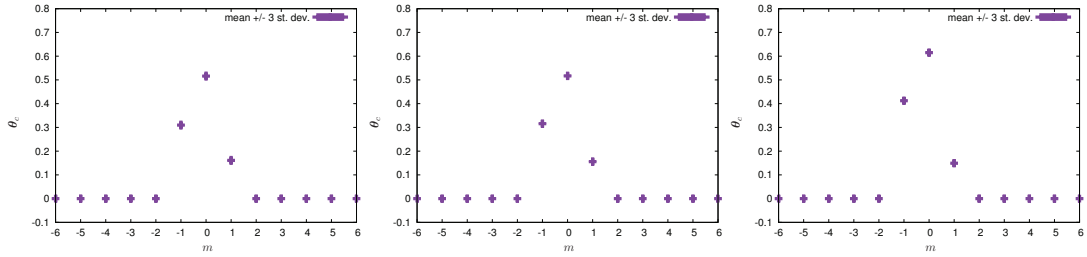


Figure 10: Evolution of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38) for $N=64$.



(a) $N=32$ (b) $N=64$ (c) $N=128$

Figure 11: Posterior estimates for $N=64$ of $\boldsymbol{\theta}_c$ associated with first-order feature functions (Eq. (3.1)).

with the size of the problem domain which is $[-1,1]$ and the CG time-step which was set to $\Delta t=1$ i.e. 400 times larger than the FG one.

At the CG level we employed $n_c=24$ variables (i.e. $n_f/n_c=100$) and as described in Section 2.3 initialized the walkers using N randomly selected initial densities and propagated them for 400 FG time-steps δt (i.e. one CG time-step Δt) in order to generate N training data. Given the availability of closed-form expressions for the evolution law of the walker-density $\rho(y,t)$ (in the limit of scale separation), we wish to assess whether the data-driven, identified CG model resembles it. We note that due to the softmax transform (Eq. (2.12)) such a comparison is not straightforward (we write the CG evolution law for $\boldsymbol{X}_t$ and not for the density $\rho(y,t)$). In this investigation we employed the same first- and second-order feature function (Eqs. (3.1), (3.2)) for $M=6$. This gave rise to 182 features in total and the same number of coefficients $\boldsymbol{\theta}_c$ which emphasizes the need for potent model selection tools.

Fig. 10 depicts the evolution of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38) for $N=64$ training data sequences. Fig. 11 depicts the posterior mean and standard deviation of the $\boldsymbol{\theta}_c$ associated with first-order feature functions for various values of m (Eq. (3.1)) and for three different data-sizes $N=32$, 64, 128. We note that the only

active ones are for $m=0$ and $m=\pm 1$ which effectively correspond to a finite difference approximation of spatial derivatives up to order 2. This is consistent with finite difference schemes of the advection-diffusion equation. Furthermore the asymmetry of the values for $m=\pm 1$ is consistent with the advective term of the equation. The aforementioned characteristics are identified correctly even with $N=32$ and, as expected, the posterior uncertainty reduces as N increases. Parameters $\boldsymbol{\theta}_c$ associated with second-order features were all deactivated by the algorithm and are therefore omitted in the illustrations.

In Fig. 12, we compare the evolution of the walker density at various future times with the predictive posterior estimates (Section 2.5) of our model. One observes that despite the fluctuations at the fine-scale, the CG model learned from $N=64$ training data is able to capture the basic trend of the density profile and that the 95% credible interval almost always envelops the ground truth. This is true as shown even for future times up to $50\Delta t$ (i.e. $20000\delta t$) even though the training data consisted of a single Δt. In Fig. 13 we perform the same comparison but for a 4-times finer binning. As in the synthetic example previously, very good predictive accuracy is attained despite the wider credible intervals which can be attributed to the more detailed character of the observable of interest.

Finally, in Fig. 14 we depict predictive estimates pertaining to second-order statistics, namely the probability of finding two walkers, simultaneously, at two bins (k_1,k_2) (2-bin probability as in the synthetic example). The comparison of the reference values (i.e. those obtained by simulating the FG model) with the probabilistic predictive estimates of the proposed model is carried out for some indicative values of (k_1,k_2), for different time instants in the future (up to $40\Delta t=16000\delta t$), and for two different numbers of bins i.e. 24 (left column) and 96 (right column). Accurate predictive estimates are observed and, as expected, the credible intervals are wider when a finer binning is employed.

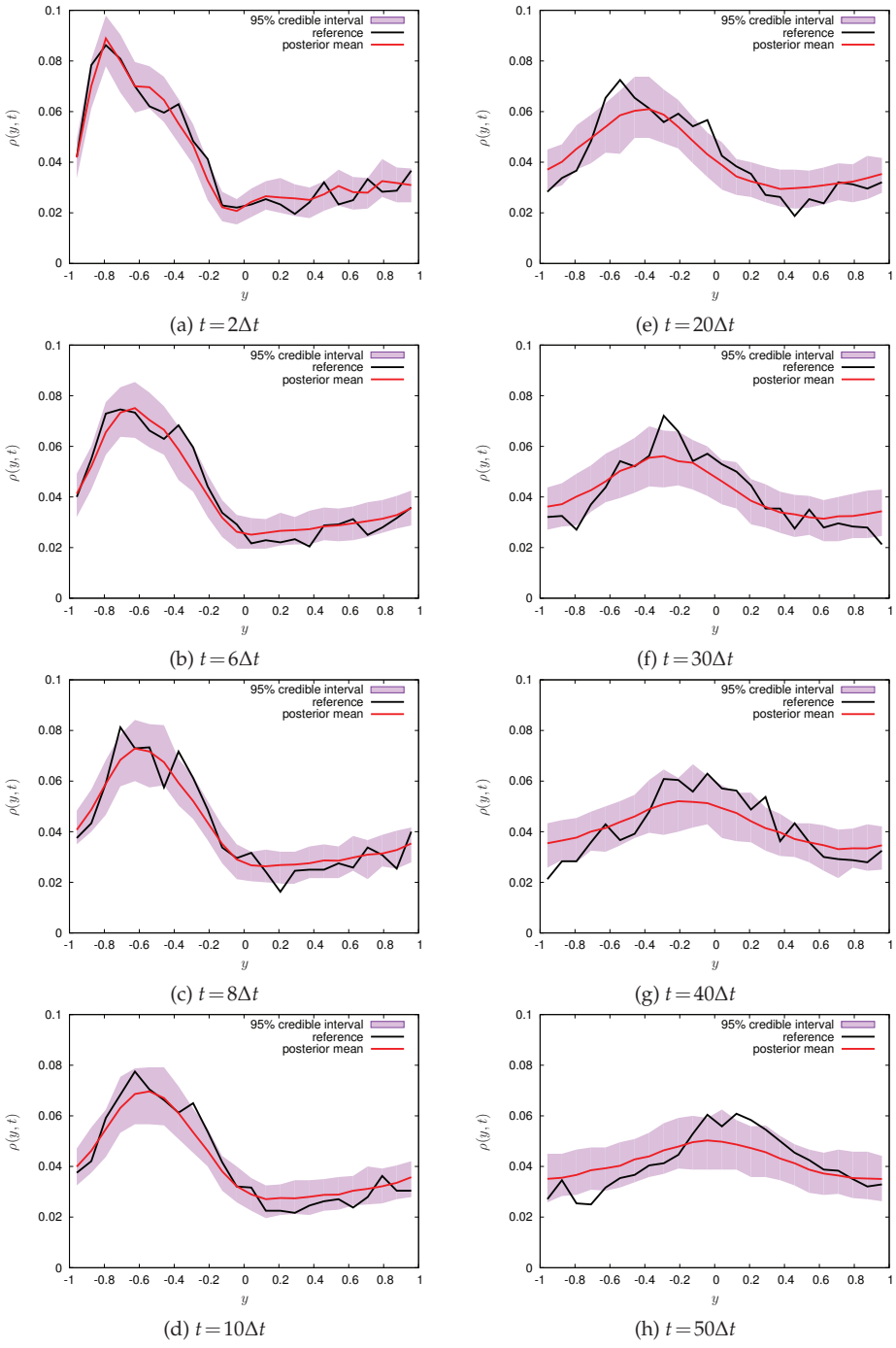
3.3 Inviscid Burgers' equation

The third and last example investigated consists at the fine-scale of $n_f=2400$ walkers[‖] which perform *interactive* random walks i.e. the jump performed at each fine-scale time-step $\delta t=2.5\times 10^{-3}$ depends on the positions of the other walkers. In particular we adopted interactions as described in [20, 90, 91] so as, in the limit, the walker density $\rho(t,y)$ follows the inviscid Burgers' equation:

$$\frac{\partial \rho}{\partial t}+\rho\frac{\partial \rho}{\partial y}=0. \tag{3.3}$$

At the CG level we employed $n_c=24$ variables (i.e. $n_f/n_c=100$) and as described in Section 2.3 initialized the walkers using N randomly selected initial densities and propagated them for 400 FG time-steps δt (i.e. one CG time-step Δt) in order to generate N

[‖] As in the previous example, we employed $\delta y=3.875\times 10^{-3}$, $\delta t=2.5\times 10^{-3}$. The latter two numbers should be compared with the size of the problem domain which is $[-1,1]$ and the CG time-step which was set to $\Delta t=1$ i.e. 400 times larger than the FG one.



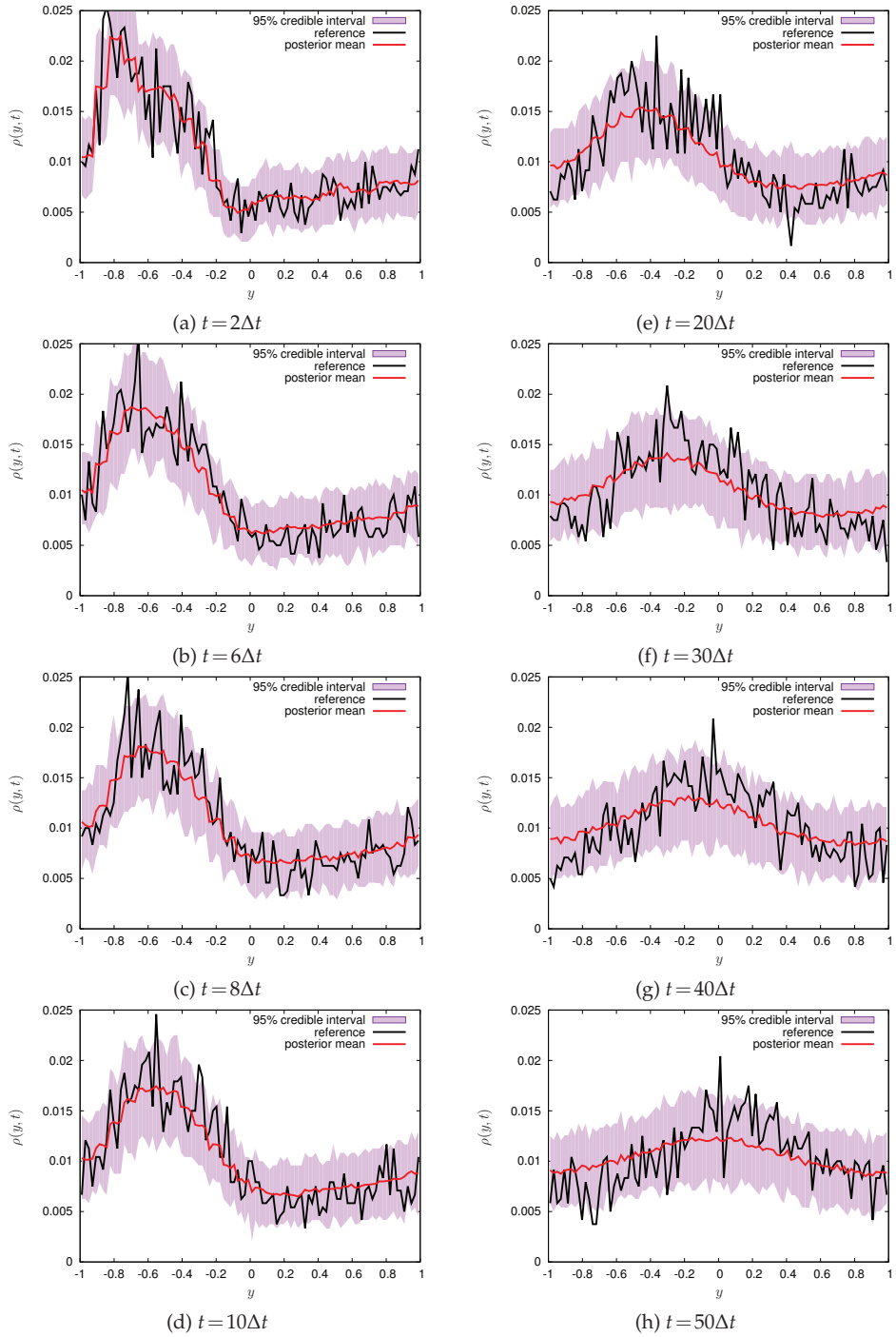
(a) $t=2\Delta t$ (e) $t=20\Delta t$

(b) $t=6\Delta t$ (f) $t=30\Delta t$

(c) $t=8\Delta t$ (g) $t=40\Delta t$

(d) $t=10\Delta t$ (h) $t=50\Delta t$

Figure 12: Predictive estimates of walker density for 24 bins (same as in CG evolution) for various future times. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=64$)



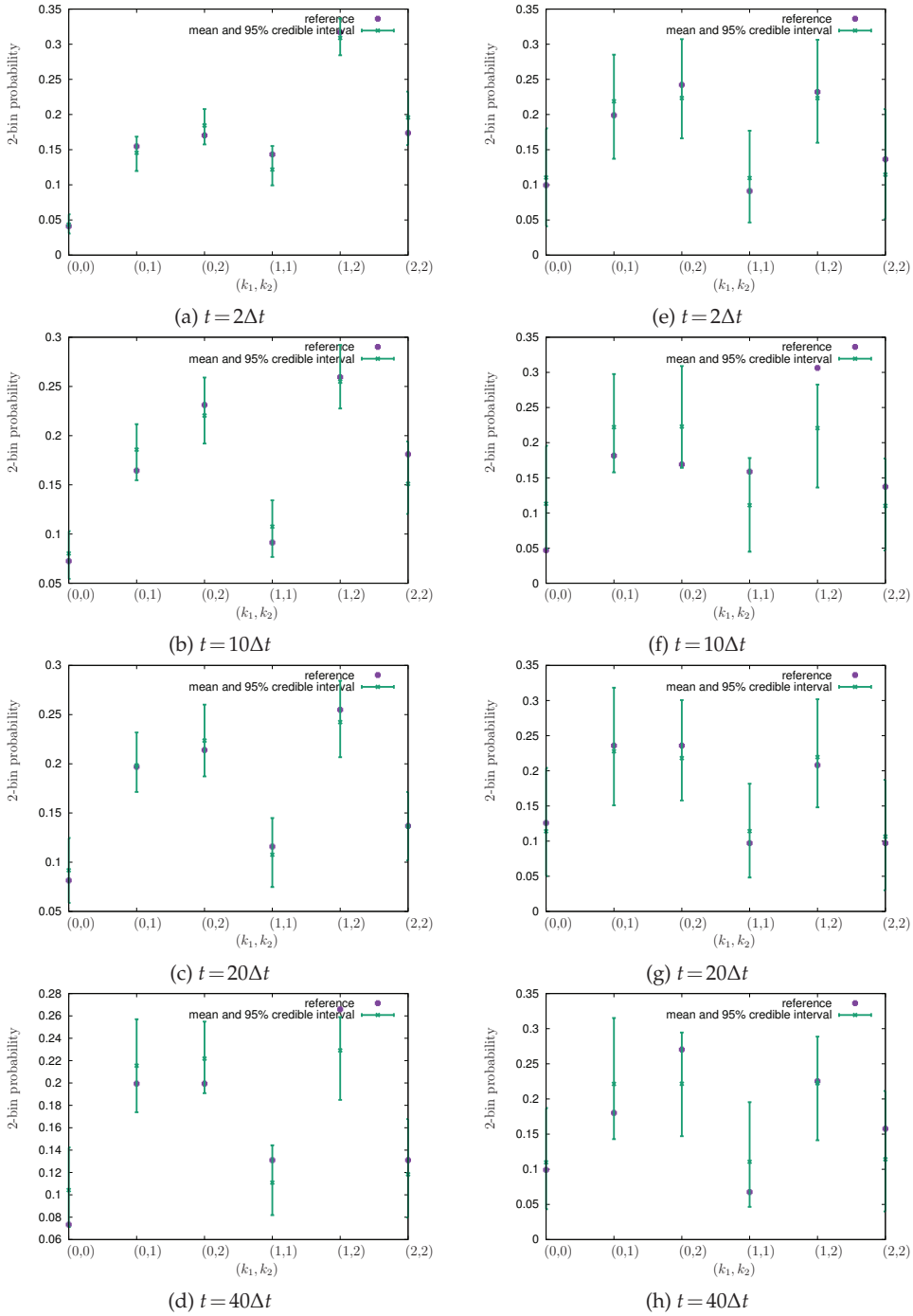
(a) $t=2\Delta t$ (e) $t=20\Delta t$

(b) $t=6\Delta t$ (f) $t=30\Delta t$

(c) $t=8\Delta t$ (g) $t=40\Delta t$

(d) $t=10\Delta t$ (h) $t=50\Delta t$

Figure 13: Predictive estimates of walker density *for* 96 *bins* (4 times greater than the 24 used in CG evolution), for various times in the future. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=64$)



(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=10\Delta t$ (f) $t=10\Delta t$

(c) $t=20\Delta t$ (g) $t=20\Delta t$

(d) $t=40\Delta t$ (h) $t=40\Delta t$

Figure 14: Predictive estimates of 2-bin probability for various times in the future i.e. $t=2\Delta t,10\Delta t,20\Delta t,40\Delta t$. The left column corresponds to the original binning, whereas the right column to 4 times finer binning. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=64$)

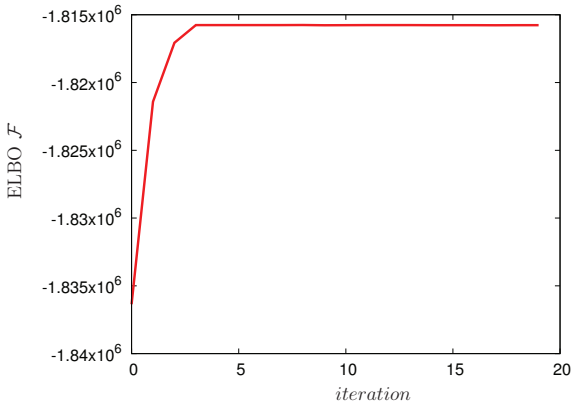


Figure 15: Evolution of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38) for $N=128$.

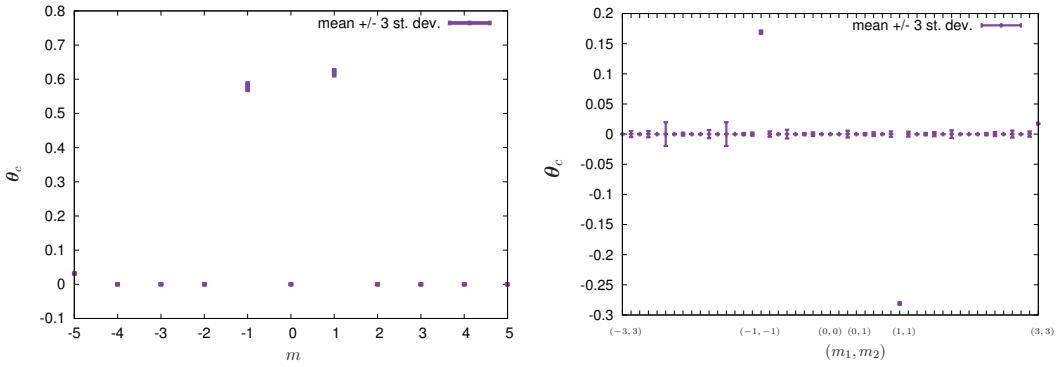


Figure 16: Posterior estimates for $N=128$ of $\boldsymbol{\theta}_c$ associated with first- (Eq. (3.1)) and second-order (Eq. (3.2)) feature functions. We plot only a subset of the coefficients of the second-order features for clarity. The omitted coefficients were practically zero (deactivated).

training data. In this investigation we employed the same first- and second-order feature function (Eqs. (3.1), (3.2)) for $M=5$. This gave rise to 132 features in total and the same number of coefficients $\boldsymbol{\theta}_c$.

Fig. 15 depicts the evolution of the ELBO $\mathcal{F}$ as estimated at each iteration based on Eq. (2.38) for $N=64$ training data sequences. Fig. 16 depicts the posterior mean and standard deviation of the $\boldsymbol{\theta}_c$ associated with the aforementioned feature functions for various values of m (Eq. (3.1)) and for $N=128$. We note that the notably active ones are for $m=\pm 1$ (first-order) and for $(-1,-1)$ and $(1,1)$ (second-order). This could be compared with the Lax-Friedrichs discretization scheme (with a spatial step δy and time-step δt) of Eq. (3.3) which is:

$$\rho_j^{t+1}=\frac{1}{2}\left(\rho_{j-1}^t+\rho_{j+1}^t\right)-\frac{\delta y}{4\delta t}\left((\rho_{j+1}^t)^2-(\rho_{j-1}^t)^2\right). \tag{3.4}$$

While, as previously noted, an immediate comparison is not possible (due to the fact the CG evolution is written with respect to $\boldsymbol{X}$ which are transformed with the softmax

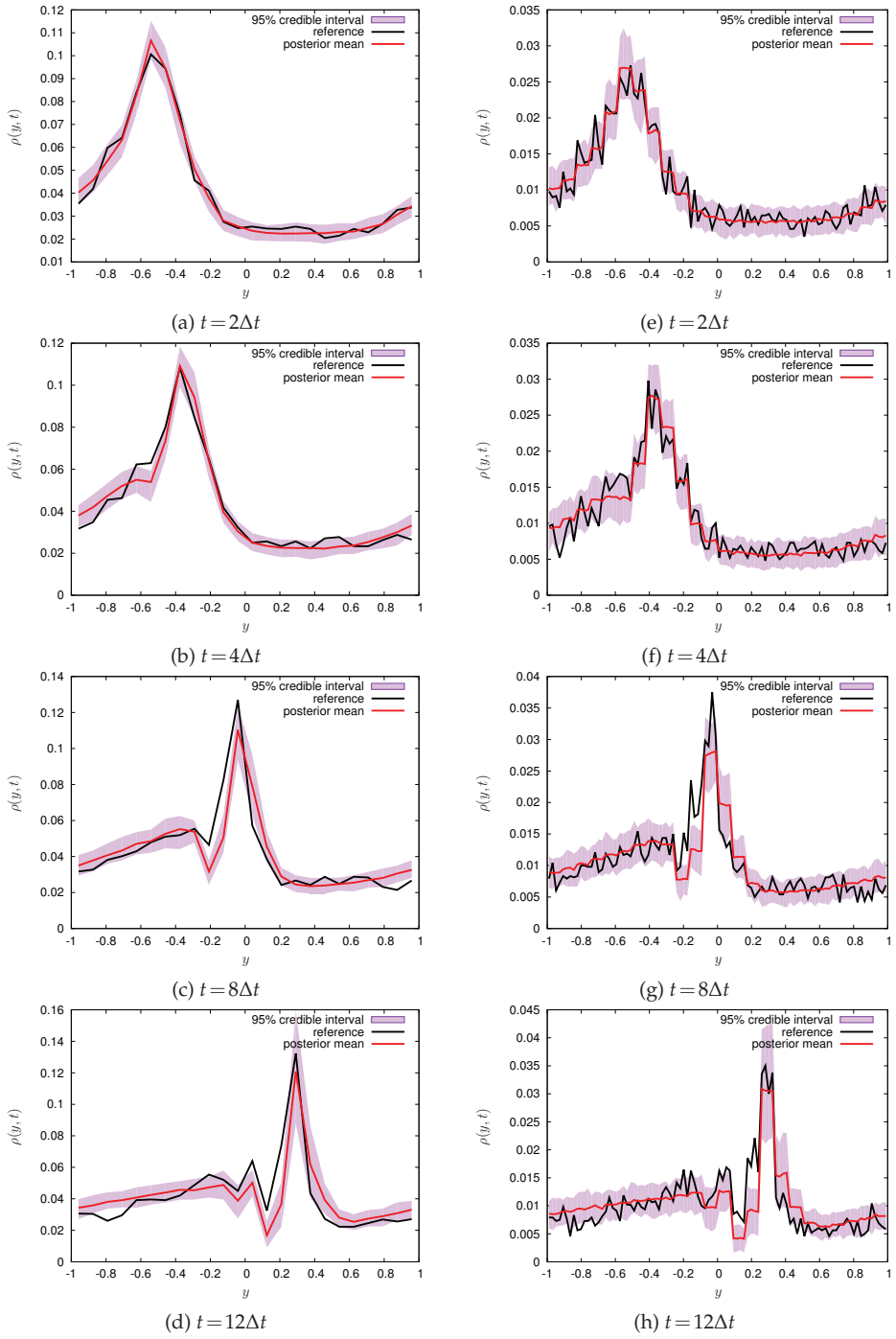
function in Eq. (2.12)), we cannot help but note that the same type of interactions are identified i.e. the same general structure for the (discretized) CG evolution law. The values of the corresponding $\boldsymbol{\theta}_c$ are not as Eq. (3.4) would suggest (e.g. the coefficients for the first order features at $m=\pm 1$ are not equal) which could be justified on the basis of the previous argument.

The most critical perhaps aspect pertains to the predictive ability of the CG model learned. In Fig. 17 we compare the reference walker density (as computed from the FG model) with the predictive estimates (posterior mean and credible intervals) for an indicative initial condition. The comparison (as in the previous example) was also performed for 4 times finer binning than that used in the CG law which indicates the ability of the model to reconstruct, albeit with uncertainty, the full fine-scale picture. As before, the predictive uncertainty is bigger when finer details are sought. In Fig. 18 we perform the same comparison, but for an initial condition that leads to a shock formation (in terms of the walker density) i.e. the walkers tend to concentrate in a (very) small region of the problem domain. As it can be seen in Fig. 18 and particularly at time $9\Delta t$ when the shock is more clearly formed, the CG model is able to follow its evolution and envelop it in the predictive estimates. Interestingly the results obtained suggest the possibility of a shock at an alternative position ($\approx y=0$) which warrants further investigation. Finally and for the same initial conditions as in Fig. 18, we depict in Fig. 19 predictive estimates of the 2-bin probability at various time instants. As in the previous examples, good agreement is observed.

4 Conclusions

We presented a probabilistic, generative probabilistic model for the automated construction of predictive, coarse-grained models from data generated by controlled, computational simulations of fine-grained models. Rather than trying to project the fine-scale state variables and governing equations to a reduced description/model, we follow the opposite route according to which the coarse states give rise to the fine-scale ones. In particular we employ a state-space model that consists of two probability densities, the first describing the evolution of the coarse state variables and the second the map from the coarse description to the fine. We adopt a Bayesian perspective in order to infer the hidden, coarse states corresponding to the data and calibrate the parameters of the model. As a result the uncertainty arising from information loss as well as the finiteness of the data available, can be quantified. More importantly this uncertainty can be propagated in order to produce probabilistic predictive estimates of the *full* fine-scale picture at any future time. Hence observables not contained in the coarse description can be predicted and the certainty in these predictions can be quantified.

We demonstrated the efficacy of the proposed model for high-dimensional systems of random walkers which, by adjusting the type of interactions, can give rise to a wide range of collective behaviors. In all cases, with small to modest data, very good predictive



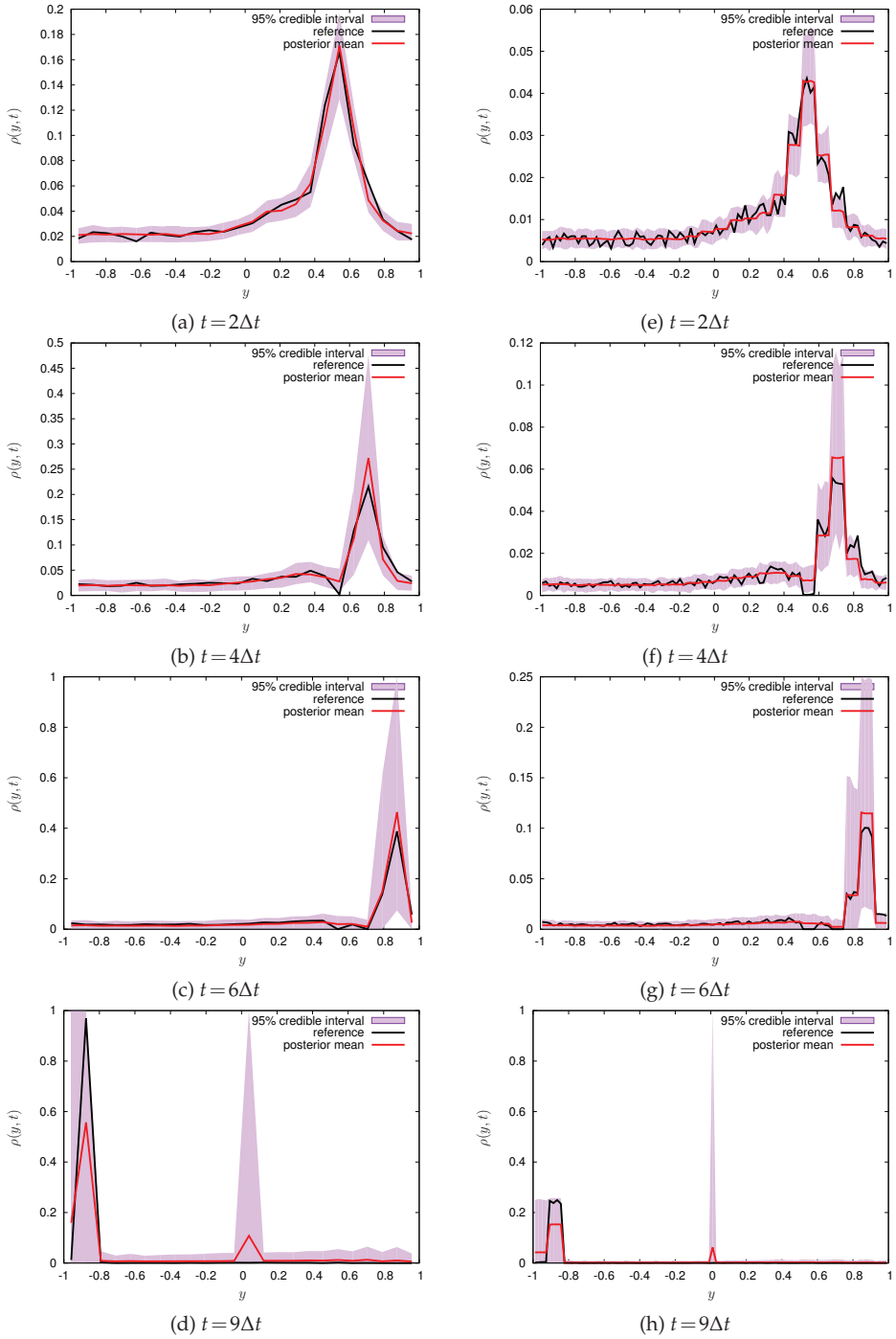
(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=4\Delta t$ (f) $t=4\Delta t$

(c) $t=8\Delta t$ (g) $t=8\Delta t$

(d) $t=12\Delta t$ (h) $t=12\Delta t$

Figure 17: Predictive estimates of walker density for various future times. The left column corresponds to the original binning, whereas the right column to 4 times finer binning. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=128$)



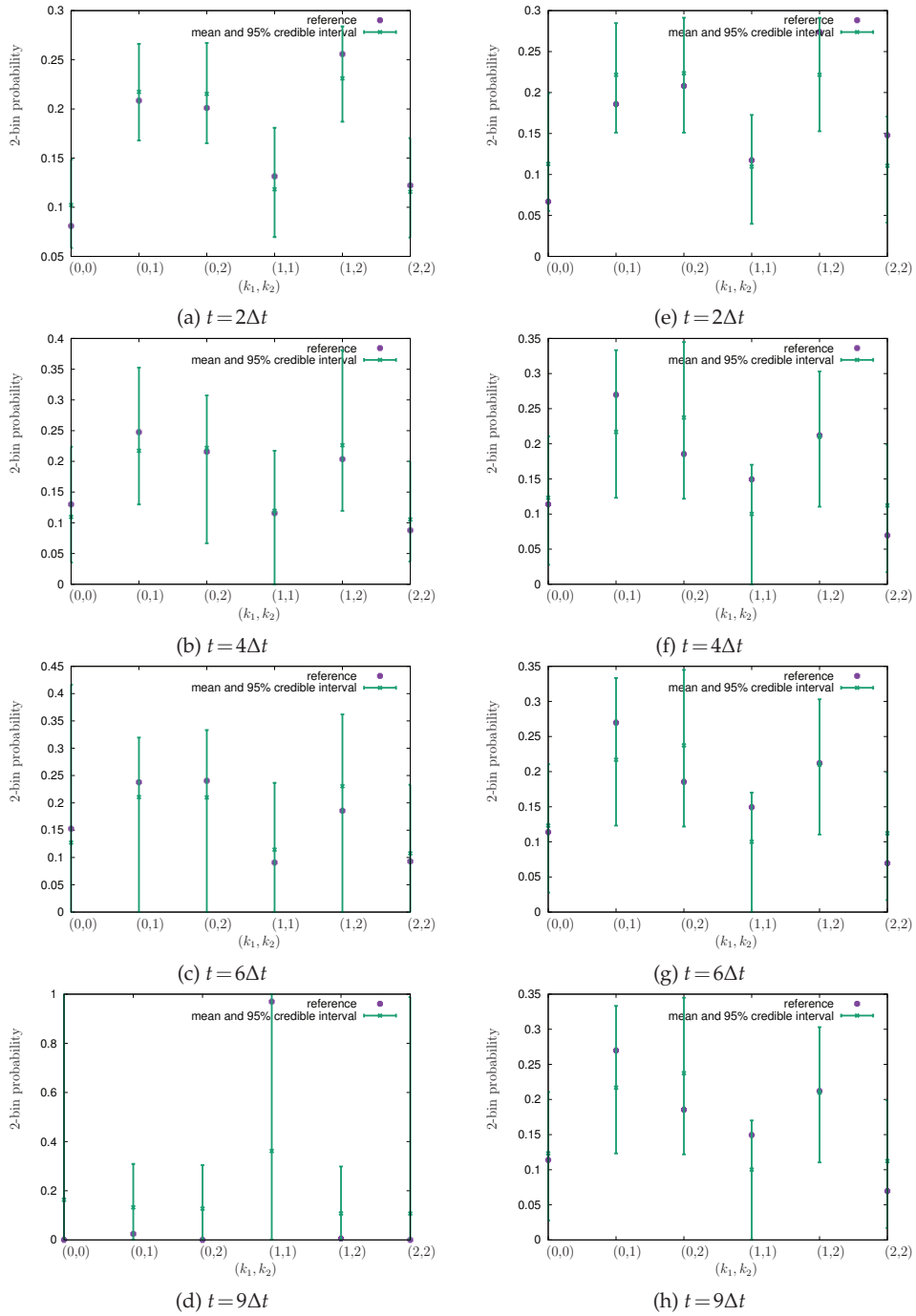
(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=4\Delta t$ (f) $t=4\Delta t$

(c) $t=6\Delta t$ (g) $t=6\Delta t$

(d) $t=9\Delta t$ (h) $t=9\Delta t$

Figure 18: Predictive estimates of walker density for various future times. The left column corresponds to the original binning, whereas the right column to 4 times finer binning. The reference density profile was computed by simulating the FG model of walkers using the FG time step δt. ($N=128$)



(a) $t=2\Delta t$ (e) $t=2\Delta t$

(b) $t=4\Delta t$ (f) $t=4\Delta t$

(c) $t=6\Delta t$ (g) $t=6\Delta t$

(d) $t=9\Delta t$ (h) $t=9\Delta t$

Figure 19: Predictive estimates of 2-bin probability for various times in the future and for the same initial conditions used in Fig. 18. The left column corresponds to the original binning, whereas the right column to 4 times finer binning. The reference 2-bin probability values were computed by simulating the FG model of walkers using the FG time step δt. ($N=64$)

estimates have been obtained even in cases where the mass density of the walkers forms shocks. An important component of the overall formulation is the discovery of salient features in the coarse model's evolution law. In the absence of prior information, this poses a formidable model selection problem which we address by employing sparsity-inducing, hierarchical priors. As demonstrated in the numerical illustrations, despite the availability of hundreds of feature functions which encode a multitude of interaction types and orders at the coarse scale, the proposed model is able to prune most of them and reveal a small subset that encodes the appropriate evolution characteristics.

Perhaps the most critical question in coarse-graining endeavors in physical modeling, pertains to the enforcement of physical constraints. In the context of the models examined, these reduce to the conservation of mass which we were able to incorporate a priori using the softmax transformation. When the coarse-grained description adopted, corresponds to continuum scales, it is well-known that additional such constraints are available from (continuum) thermodynamics in the form of conservation laws or the entropy inequality. Incorporating such constraints/invariances is of primary importance not only because it ensures physically-meaningful predictions but also because it has the potential of significantly reducing the amount of training data needed. Particularly since the data are generated from expensive, fine-scale simulations which unavoidably explore a (small) portion of the configuration space. An automated, general framework for enforcing such constraints is currently missing from the vocabulary of probabilistic modeling, and in our opinion, represents the most important challenge along this modeling direction.

Appendix A

In this section we derive the expression of the ELBO $\mathcal{F}$ based on the original formula in Eq. (2.27) and the optimal approximating densities q for the latent variables of the model as presented in Section 2.4. We note that (up to constants):

$$
\begin{aligned}
\mathcal{F} =& \int q(\boldsymbol{X}_{\Delta t}^{(1:N)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}, \boldsymbol{v}) \log p(\boldsymbol{x}_{\Delta t}^{(1:N)} | \boldsymbol{X}_{\Delta t}^{(1:N)}) \, d\boldsymbol{X}_{\Delta t}^{(1:N)} \, d\boldsymbol{\theta}_c \, d\boldsymbol{\tau} \, d\boldsymbol{v} \\
&+ \int q(\boldsymbol{X}_{\Delta t}^{(1:N)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}, \boldsymbol{v}) \log \frac{p(\boldsymbol{X}_{\Delta t}^{(1:N)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}, \boldsymbol{v})}{q(\boldsymbol{X}_{\Delta t}^{(1:N)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}, \boldsymbol{v})} \, d\boldsymbol{X}_{\Delta t}^{(1:N)} \, d\boldsymbol{\theta}_c \, d\boldsymbol{\tau} \, d\boldsymbol{v} \\
=& \sum_{i=1}^{N} < \log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)} | \boldsymbol{X}_{\Delta t}^{(i)}) >_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})} \qquad \text{(from Eq. (2.23))} \\
&+ \sum_{i=1}^{N} < \log p_{cf}(\boldsymbol{X}_{\Delta t}^{(i)} | \boldsymbol{X}_{0}^{(i)}, \boldsymbol{\theta}_c, \boldsymbol{v}) >_{q(\boldsymbol{X}_{\Delta t}^{(i)}) q(\boldsymbol{\theta}_c) q(\boldsymbol{v})} \qquad \text{(from Eqs. (2.24), (2.25))} \\
&+ < \log p(\boldsymbol{\theta}_c | \boldsymbol{\tau}) >_{q(\boldsymbol{\theta}_c) q(\boldsymbol{\tau})} + < \log p(\boldsymbol{\tau}) >_{q(\boldsymbol{\tau}} + < \log p(\boldsymbol{v}) >_{q(\boldsymbol{v})} \qquad \text{(from Eq. (2.24))} \\
&- \sum_{i=1}^{N} < q(\boldsymbol{X}_{\Delta t}^{(i)}) >_{q(\boldsymbol{X}_{\Delta t}^{(i)})} - < \log q(\boldsymbol{\theta}_c) >_{q(\boldsymbol{\theta}_c)}
\end{aligned}
$$

$$
\begin{aligned}
&-<\log q(\boldsymbol{\tau})>_{q(\boldsymbol{\tau})}-<\log q(\boldsymbol{v})>_{q(\boldsymbol{v})} \qquad \text{(from Eq. (2.28))}\\
=&\sum_{i=1}^{N}<\log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})}\\
&+\frac{N}{2}\sum_{j=1}^{n_c}<log v_j>_{q(\boldsymbol{v})}\\
&-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_c}<v_j>_{q(\boldsymbol{v})}<(X_{(k+1)\Delta t,\,j}^{(i)}-\boldsymbol{\theta}_c^T\boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0^{(i)}))^2>_{q(\boldsymbol{X}_{\Delta t}^{(i)})q(\boldsymbol{\theta}_c)} \qquad \text{(from Eq. (2.17))}\\
&+\frac{1}{2}\sum_{l=1}^{L}<\log\tau_l>_{q(\boldsymbol{\tau})}-\frac{1}{2}<\tau_l>_{q(\boldsymbol{\tau})}<\theta_{c,l}^2>_{q(\boldsymbol{\theta}_c)} \qquad \text{(from Eq. (2.19))}\\
&+(\alpha_0-1)\sum_{l=1}^{L}<\log\tau_l>_{q(\boldsymbol{\tau})}-\beta_0\sum_{l=1}^{L}<\tau_l>_{q(\boldsymbol{\tau})} \qquad \text{(from Eq. (2.20))}\\
&+(\gamma_0-1)\sum_{j=1}^{n_c}<\log v_j>_{q(\boldsymbol{v})}-\zeta_0\sum_{j=1}^{n_c}<v_j>_{q(\boldsymbol{v})} \qquad \text{(from Eq. (2.21))}\\
&+\frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{S}_i| \qquad \text{(from Eq. (2.37))}\\
&+\frac{1}{2}\log|\boldsymbol{S}_{\boldsymbol{\theta}}| \qquad \text{(from Eq. (2.33))}\\
&-\sum_{l}^{L}(\alpha_l-1)<\log\tau_l>_{q(\boldsymbol{\tau})}+\sum_{l=1}^{L}\beta_l<\tau_l>_{q(\boldsymbol{\tau})}+\sum_{l=1}^{L}\log Z(\alpha_l,\beta_l) \qquad \text{(from Eq. (2.34))}\\
&-\sum_{j=1}^{n_c}(\gamma_j-1)<\log v_j>_{q(\boldsymbol{v})}+\zeta_0\sum_{j=1}^{n_c}<v_j>_{q(\boldsymbol{v})}+\sum_{j=1}^{n_c}\log Z(\gamma_j,\zeta_j) \qquad \text{(from Eq. (2.35))},
\end{aligned}
\tag{A.1}
$$

where $Z(\alpha,\beta)=\frac{\Gamma(\alpha)}{\beta^{\alpha}}$ is the normalization constant of the Gamma distribution with parameters α,β. Using the forms of optimal q's as presented in Eqs. (2.33), (2.34) and (2.35), we can establish that:

$$
\begin{aligned}
\mathcal{F}=&\sum_{i=1}^{N}<\log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})}\\
&+\frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{S}_i|+\frac{1}{2}\log|\boldsymbol{S}_{\boldsymbol{\theta}}|\\
&+\sum_{l=1}^{L}\log Z(\alpha_l,\beta_l)+\sum_{j=1}^{n_c}\log Z(\gamma_j,\zeta_j),
\end{aligned}
\tag{A.2}
$$

which based on the fact that α_j and γ_j are constant (given N), can be further simplified to:

$$\mathcal{F}=\sum_{i=1}^{N}<\log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(1:N)})}+\frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{S}_i|+\frac{1}{2}\log|\boldsymbol{S}_{\boldsymbol{\theta}}| \\ -\sum_{l}^{L}\alpha_l\log\beta_l-\sum_{j=1}^{n_c}\gamma_j\log\zeta_j. \tag{A.3}$$

The first expectation is analytically intractable given the form of the p_{cf} in Eq. (2.13). For that purpose it is estimated with Monte Carlo. We finally note that lower bounds have been proposed for the log of the softmax function [92, 93] appearing in p_{cf} which could lead to closed-form updates for $q(\boldsymbol{X}_{\Delta t}^{(1:N)})$ at the expense of a poorer approximation to the log-evidence.

Appendix B

The present section contains further details for the computation of the gradient of $\mathcal{F}_i$ (Eq. (2.36)) with respect to the variational parameters $\boldsymbol{\mu}_i,\boldsymbol{S}_i$ in Eq. (2.37). These derivatives are used in the context of the ADAM algorithm for updating the latter parameters. We recall from Eq. (2.36) that:

$$\mathcal{F}_i=<\log p_{cf}(\boldsymbol{x}_{\Delta t}^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})}+<\log p_c(\boldsymbol{X}_{\Delta t}^{(i)}|\boldsymbol{X}_0^{(i)},\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})q(\boldsymbol{\theta}_c)q(\boldsymbol{v})} \\ -<\log q(\boldsymbol{X}_{\Delta t}^{(i)})>_{q(\boldsymbol{X}_{\Delta t}^{(i)})}. \tag{B.1}$$

In the following we *drop the super/sub-script i referring to the data point considered* in order to simplify the notations. Furthermore we employ a diagonal covariance for $\boldsymbol{S}=diag(\boldsymbol{s}_j^2)$ and denote with $\boldsymbol{s}\in\mathbb{R}^{n_c,+}$ the vector of the diagonal standard deviations. The latter two terms in Eq. (B.1) can be analytically computed. In particular (up to a constant):

$$<\log q(\boldsymbol{X}_{\Delta t})>_{q(\boldsymbol{X}_{\Delta t})}=-\frac{1}{2}\log|\boldsymbol{S}|. \tag{B.2}$$

Furthermore, from Eqs. (2.17) and (2.18) (up to a constant):

$$<\log p_c(\boldsymbol{X}_{\Delta t}|\boldsymbol{X}_0,\boldsymbol{\theta}_c,\boldsymbol{v})>_{q(\boldsymbol{X}_{\Delta t})q(\boldsymbol{\theta}_c)q(\boldsymbol{v})} \\ =-\frac{1}{2}\sum_{j=1}^{n_c}<v_j>_{q_j(v_j)}(\mu_j^2+s_j^2-2\mu_j<\boldsymbol{\theta}_c^T>_{q(\boldsymbol{\theta}_c)}\boldsymbol{\phi}^{(j)}(\boldsymbol{X}_0)). \tag{B.3}$$

Finally for the first term in Eq. (B.1) we employ the reparametrization trick [83]. This consists of the observation that for any function $f(\boldsymbol{X}_{\Delta t})$, its expectation with respect to $q(\boldsymbol{X}_{\Delta t})=\mathcal{N}(\boldsymbol{X}_{\Delta t};\boldsymbol{\mu},\boldsymbol{S})$ can be expressed as:

$$<f(\boldsymbol{X}_{\Delta t})>_{q(\boldsymbol{X}_t)}=<f(\boldsymbol{\mu}+\sqrt{\boldsymbol{S}}\boldsymbol{\epsilon})>_{q(\boldsymbol{\epsilon})}, \tag{B.4}$$

where $q(\boldsymbol{\epsilon})=\mathcal{N}(\boldsymbol{X}_t;\mathbf{0},\boldsymbol{I})$ is the standard normal. Hence,

$$\frac{\partial <f(\boldsymbol{X}_{\Delta t})>_{q(\boldsymbol{X}_{\Delta t})}}{\partial \boldsymbol{\mu}}=<\nabla_{\boldsymbol{X}_{\Delta t}} f(\boldsymbol{\mu}+\sqrt{\boldsymbol{S}}\boldsymbol{\epsilon})>_{q(\boldsymbol{\epsilon})} \tag{B.5}$$

and

$$\frac{\partial <f(\boldsymbol{X}_{\Delta t})>_{q(\boldsymbol{X}_{\Delta t})}}{\partial \boldsymbol{s}}=<\nabla_{\boldsymbol{X}_{\Delta t}} f(\boldsymbol{\mu}+\sqrt{\boldsymbol{S}}\boldsymbol{\epsilon})\otimes\boldsymbol{\epsilon}>_{q(\boldsymbol{\epsilon})}, \tag{B.6}$$

where $\otimes$ is the Hadamard (element-wise) product. The derivatives of the first term of Eq. (B.1) are computed by substituting $\log p_{cf}(\boldsymbol{x}_{\Delta t}|\boldsymbol{X}_{\Delta t})$ in the place of f above and using the functional form of p_{cf} from Eq. (2.13).

References

[1] J. Harlim and X. Li, "Parametric reduced models for the nonlinear Schr\"odinger equation," *Physical Review E*, vol. 91, p. 053306, May 2015.

[2] X. Li, "A coarse-grained molecular dynamics model for crystalline solids," *International Journal for Numerical Methods in Engineering*, vol. 83, pp. 986–997, Aug. 2010.

[3] A. Majda, C. Franzke, and D. Crommelin, "Normal Forms for Reduced Stochastic Climate Models," *PNAS*, vol. 16, no. 10, pp. 3649–3653, 2009.

[4] A. J. Majda and J. Harlim, "Physics constrained nonlinear regression models for time series," *Nonlinearity*, vol. 26, no. 1, p. 201, 2013.

[5] Z. Li, X. Bian, X. Yang, and G. E. Karniadakis, "A comparative study of coarse-graining methods for polymeric fluids: Mori-Zwanzig vs. iterative Boltzmann inversion vs. stochastic parametric optimization," *The Journal of Chemical Physics*, vol. 145, p. 044102, July 2016.

[6] S. A. Thomas, D. J. B. Lloyd, and A. C. Skeldon, "Equation-free analysis of agent-based models and systematic parameter determination," *Physica A: Statistical Mechanics and its Applications*, vol. 464, pp. 27–53, Dec. 2016.

[7] D. Givon, R. Kupferman, and A. Stuart, "Extracting Macroscopic Dynamics: Model Problems and Algorithms," *Nonlinearity*, 2004.

[8] I. Horenko, F. Noe, C. Hartmann, and C. Schtte, "Data-based parameter estimation of generalized multidimensional Langevin processes," *Physical Review E*, vol. 78, p. 016706, 2007.

[9] P. Espanol and I. Zuniga, "Obtaining fully dynamic coarse-grained models from MD," *Physical Chemistry Chemical Physics*, vol. 13, pp. 10538–10545, May 2011.

[10] J. Wan and N. Zabaras, "A probabilistic graphical model approach to stochastic multiscale partial differential equations," *Journal of Computational Physics*, vol. 250, pp. 477–510, Oct. 2013.

[11] D. Kondrashov, M. D. Chekroun, and M. Ghil, "Data-driven non-Markovian closure models," *Physica D: Nonlinear Phenomena*, vol. 297, pp. 33–55, Mar. 2015.

[12] V. Harmandaris, E. Kalligiannaki, M. A. Katsoulakis, and P. Plech, "Path-space variational inference for non-equilibrium coarse- grained systems," *Journal of Computational Physics*, Mar. 2016.

[13] H. Lei, N. A. Baker, and X. Li, "Data-driven parameterization of the generalized Langevin equation," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 14183–14188, Dec. 2016.

[14] M. Schoeberl, N. Zabaras, and P.-S. Koutsourelakis, "Predictive coarse-graining," *Journal of Computational Physics*, vol. 333, pp. 49–77, Mar. 2017.

[15] M. Crosskey and M. Maggioni, "ATLAS: A Geometric Approach to Learning High-Dimensional Stochastic Systems Near Manifolds," *Multiscale Modeling & Simulation*, vol. 15, pp. 110–156, Jan. 2017.
[16] Q. Yang, C. A. Sing-Long, and E. J. Reed, "Learning reduced kinetic Monte Carlo models of complex chemistry from molecular dynamics," *Chemical Science*, June 2017.
[17] M. Raissi and G. E. Karniadakis, "Hidden physics models: Machine learning of nonlinear partial differential equations," *Journal of Computational Physics*, vol. 357, pp. 125–141, Mar. 2018.
[18] M. Raissi, "Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations," *arXiv:1801.06637 [cs, math, stat]*, Jan. 2018. arXiv: 1801.06637.
[19] T. Hey, S. Tansley, K. M. Tolle, and others, *The fourth paradigm: data-intensive scientific discovery*, vol. 1. Microsoft research Redmond, WA, 2009.
[20] J. Li, P. G. Kevrekidis, C. W. Gear, and I. G. Kevrekidis, "Deciding the Nature of the Coarse Equation Through Microscopic Simulations: The Baby-Bathwater Scheme," *SIAM Rev.*, vol. 49, pp. 469–487, July 2007.
[21] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science (New York, N.Y.)*, vol. 324, pp. 81–85, Apr. 2009.
[22] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 3932–3937, Apr. 2016.
[23] J. Han, L. Zhang, R. Car, and E. Weinan, "Deep Potential: A General Representation of a Many-Body Potential Energy Surface," *COMMUNICATIONS IN COMPUTATIONAL PHYSICS*, vol. 23, pp. 629–639, Mar. 2018.
[24] H. Grabert, *Projection operator techniques in nonequilibrium statistical mechanics*. Springer-Verlag, 1982.
[25] M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," *Annual review of physical chemistry*, vol. 64, pp. 295–316, 2013.
[26] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *APPLIED AND COMPUTATIONAL HARMONIC ANALYSIS*, vol. 21, no. 1, pp. 113 – 127, 2006.
[27] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, "Heterogeneous multiscale methods: a review," *Commun. Comput. Phys*, vol. 2, no. 3, pp. 367–450, 2007.
[28] I. Kevrekidis, C. Gear, J. Hyman, P. Kevrekidis, O. Runborg, and K. Theodoropoulos, "Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks," *Communications in Mathematical Sciences*, vol. 1, no. 4, pp. 715–762, 2003.
[29] H. Mori, "Transport, Collective Motion, and Brownian Motion," *Progress of Theoretical Physics*, vol. 33, pp. 423–455, Mar. 1965.
[30] R. Zwanzig, "Nonlinear generalized Langevin equations," *Journal of Statistical Physics*, vol. 9, pp. 215–220, Nov. 1973.
[31] A. Chorin, O. Hald, and R. Kupferman, "Optimal prediction and the Mori-Zwanzig representation of irreversible processes," *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, vol. 97, no. 7, pp. 2968 – 2973, 2000.
[32] A. Chorin and P. Stinis, "Problem reduction, renormalization, and memory," *Comm. Appl. Math. Comp. Sc.*, vol. 1, pp. 1–27, 2005.
[33] P. Stinis, "Renormalized Mori-Zwanzig-reduced models for systems without scale separation," *Proceedings of the Royal Society of London Series A*, vol. 471, pp. 20140446–20140446, Feb.

2015.
[34] F. Legoll and T. Lelivre, "Effective dynamics using conditional expectations," *Nonlinearity*, vol. 23, no. 9, p. 2131, 2010.
[35] E. Darve, J. Solomon, and A. Kia, "Computing generalized Langevin equations and generalized FokkerPlanck equations," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 10884–10889, July 2009.
[36] C. Hijon, P. Espanol, E. Vanden-Eijnden, and R. Delgado-Buscalioni, "Mori-Zwanzig formalism as a practical computational tool," *Faraday discussions*, vol. 144, pp. 301–322; discussion 323–345, 467–481, 2010.
[37] B. D. Coleman and M. E. Gurtin, "Thermodynamics with Internal State Variables," *The Journal of Chemical Physics*, vol. 47, pp. 597–613, July 1967.
[38] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science*. New York: Springer, 3rd ed. 2013 edition ed., May 2013.
[39] R. Erban, T. Frewen, X. Wang, T. Elston, R. Coifman, B. Nadler, and I. Kevrekidis, "Variable-free exploration of stochastic models: A gene regulatory network example," *JOURNAL OF CHEMICAL PHYSICS*, vol. 126, no. 15, p. 155103, 2007.
[40] M. A. Katsoulakis and J. Trashorras, "Information loss in coarse-graining of stochastic particle dynamics," *Journal of statistical physics*, vol. 122, no. 1, pp. 115–135, 2006.
[41] D. V. Lindley, "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics*, vol. 27, pp. 986–1005, Dec. 1956.
[42] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*. Springer-Verlag, 2005.
[43] Z. Ghahramani, "Unsupervised Learning," in *Advanced Lectures on Machine Learning LNAI 3176* (O. Bousquet, G. Raetsch, and U. von Luxburg, eds.), Springer-Verlag, 2004.
[44] D. Durstewitz, "A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements," *PLOS Computational Biology*, vol. 13, p. e1005542, June 2017.
[45] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
[46] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, "Predicting Catastrophes in Nonlinear Dynamical Systems by Compressive Sensing," *Physical Review Letters*, vol. 106, p. 154101, Apr. 2011.
[47] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher, "Sparse dynamics for partial differential equations," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 6634–6639, Apr. 2013.
[48] J. L. Proctor, S. L. Brunton, B. W. Brunton, and J. N. Kutz, "Exploiting sparsity and equation-free architectures in complex systems," *The European Physical Journal Special Topics*, vol. 223, pp. 2665–2684, Dec. 2014.
[49] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, June 1996.
[50] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, pp. 3311–3325, Dec. 1997.
[51] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006* (B. Schoelkopf, J. C. Platt, and T. Hoffman, eds.), pp. 801–808, MIT Press, 2006.
[52] I. J. Goodfellow, A. Courville, and Y. Bengio, "Spike-and-slab sparse coding for unsupervised feature discovery," *arXiv preprint arXiv:1201.3382*, 2012.
[53] M. Chalk, O. Marre, and G. Tkacik, "Relevant sparse codes with variational information

bottleneck," *arXiv:1605.07332 [stat]*, May 2016.
[54] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.
[55] D. P. Wipf and S. S. Nagarajan, "A New View of Automatic Relevance Determination," in *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 1625–1632, Curran Associates, Inc., 2008.
[56] D. Mackay, "Probable Networks and Plausible Predictions - a Review of Practical Bayesian Methods for Supervised Neural Networks," *Network-Computation in Neural Systems*, vol. 6, pp. 469–505, Aug. 1995.
[57] J. W. Paisley, D. M. Blei, and M. I. Jordan, "Variational Bayesian Inference with Stochastic Search," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.
[58] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic Variational Inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, May 2013.
[59] B. C. Daniels and I. Nemenman, "Automated adaptive inference of phenomenological dynamical models," *Nature Communications*, vol. 6, p. ncomms9133, Aug. 2015.
[60] Z. Y. Wan and T. P. Sapsis, "Reduced-space Gaussian Process Regression for data-driven probabilistic forecast of chaotic dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 345, pp. 40–55, Apr. 2017.
[61] H. Schaeffer, "Learning partial differential equations via data discovery and sparse optimization," *Proc. R. Soc. A*, vol. 473, p. 20160446, Jan. 2017.
[62] G. N. Milstein, *Numerical integration of stochastic differential equations*, vol. 313. Springer Science & Business Media, 1994.
[63] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.
[64] M. A. Katsoulakis and P. Plech, "Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems," *The Journal of Chemical Physics*, vol. 139, p. 074115, Aug. 2013.
[65] M. E. Gurtin, E. Fried, and L. Anand, *The Mechanics and Thermodynamics of Continua*. Cambridge: Cambridge University Press, reprint edition ed., Dec. 2013.
[66] D. Aldous, "Exchangeability and related topics," *Ecole d'Ete de Probabilites de Saint-Flour XII*, pp. 1–198, 1985.
[67] M. I. Jordan, "Bayesian nonparametric learning: Expressive priors for intelligent systems," *Heuristics, probability and causality: A tribute to Judea Pearl*, vol. 11, pp. 167–185, 2010.
[68] C. Bishop, "Latent variable models," in *Learning in Graphical Models* (M. I. Jordan, ed.), pp. 371–403, MIT Press, 1999.
[69] M. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society B*, vol. 61, pp. 611–622, 1999.
[70] Z. Long, Y. Lu, X. Ma, and B. Dong, "Pde-net: Learning pdes from data," *CoRR*, vol. abs/1710.09668, 2017.
[71] J. F. Rudzinski and W. G. Noid, "Coarse-graining entropy, forces, and structures," *The Journal of Chemical Physics*, vol. 135, p. 214101, 2011.
[72] W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," *The Journal of Chemical Physics*, vol. 139, p. 090901, Sept. 2013.
[73] I. Bilionis and P. S. Koutsourelakis, "Free energy computations by minimization of KullbackLeibler divergence: An efficient adaptive biasing potential method for sparse representations," *Journal of Computational Physics*, vol. 231, no. 9, pp. 3849 – 3870, 2012.
[74] M. A. T. Figueiredo, "Adaptive Sparseness for Supervised Learning," *IEEE Trans. Pattern*

Anal. Mach. Intell., vol. 25, no. 9, pp. 1150–1159, 2003.
[75] D. J. C. MacKay and R. M. Neal, "Automatic relevance determination for neural networks," tech. rep., University of Cambridge, 1994.
[76] M. E. Tipping, "The Relevance Vector Machine," in *Advances in Neural Information Processing Systems 12* (S. A. Solla, T. K. Leen, and K.-R. Mller, eds.), pp. 652–658, MIT Press, 2000.
[77] R. Peierls, "On a Minimum Property of the Free Energy," *Physical Review*, vol. 54, pp. 918–919, Dec. 1938.
[78] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," in *Foundations and Trends in Machine Learning*, vol. 1 of *1-305*, pp. 1–305, 2008.
[79] G. Parisi, *Statistical Field Theory*. Redwood City, Calif: Addison Wesley, Jan. 1988.
[80] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 1st ed. 2006. corr. 2nd printing 2011 ed., 2007.
[81] H. Robbins and S. Monro, "A stochastic approximation method," *Annals Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.
[82] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The International Conference on Learning Representations (ICLR)*, (San Diego), 2015.
[83] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *The International Conference on Learning Representations (ICLR)*, vol. abs/1312.6114, (Banff, Alberta, Canada), 2014.
[84] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[85] R. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, pp. 355–368, Kluwer Academic Publishers, 1998.
[86] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, vol. 68, pp. 411–436, 2006.
[87] M. Titsias and M. Lazaro-Gredilla, "Doubly stochastic variational Bayes for non-conjugate inference," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.
[88] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[89] G.-H. Cottet and P. D. Koumoutsakos, *Vortex Methods: Theory and Practice*. Cambridge ; New York: Cambridge University Press, 2 edition ed., Mar. 2000.
[90] S. Roberts, "Convergence of a Random Walk Method for the Burgers Equation," *Mathematics of Computation*, vol. 52, no. 186, pp. 647–673, 1989.
[91] A. Chertock and D. Levy, "Particle Methods for Dispersive Equations," *Journal of Computational Physics*, vol. 171, pp. 708–730, Aug. 2001.
[92] G. Bouchard, "Efficient bounds for the softmax function, applications to inference in hybrid models," in *NIPS*, 2007.
[93] M. K. Titsias, "One-vs-Each Approximation to Softmax for Scalable Estimation of Probabilities," in *NIPS 29*, pp. 4161–4169, 2016.