Journal of Computational Physics 514 (2024) 113206

Contents lists available at ScienceDirect

Journal of Computational Physics

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

Automated construction of effective potential via algorithmic implicit bias

Xingjie Helen Li [a,*], Molei Tao [b]

[a] Department of Mathematics and Statistics, University of North Carolina at Charlotte, United States of America
[b] School of Mathematics, Georgia Institute of Technology, United States of America

A B S T R A C T

We introduce a novel approach for decomposing and learning every scale of a given multiscale objective function in $\mathbb{R}^d$, where $d \geqslant 1$. This approach leverages a recently demonstrated implicit bias of the optimization method of gradient descent [44], which enables the automatic generation of data that nearly follow Gibbs distribution with an effective potential at any desired scale. One application of this automated effective potential modeling is to construct reduced-order models. For instance, a deterministic surrogate Hamiltonian model can be developed to substantially soften the stiffness that bottlenecks the simulation, while maintaining the accuracy of phase portraits at the scale of interest. Similarly, a stochastic surrogate model can be constructed at a desired scale, such that both its equilibrium and out-of-equilibrium behaviors (characterized by auto-correlation function and mean path) align with those of a damped mechanical system with the original multiscale function being its potential. The robustness and efficiency of our proposed approach in multi-dimensional scenarios have been demonstrated through a series of numerical experiments. A by-product of our development is a method for anisotropic noise estimation and calibration. More precisely, Langevin model of stochastic mechanical systems may not have isotropic noise in practice, and we provide a systematic algorithm to quantify its covariance matrix without directly measuring the noise. In this case, the system may not admit closed form expression of its invariant distribution either, but with this tool, we can design friction matrix appropriately to calibrate the system so that its invariant distribution has a closed form expression of Gibbs.

1. Introduction

Physical and data sciences encompass numerous problems that involve multiple spatial and temporal scales. For example, these problems may involve the evolution of a meta-particle within a multiscale energy landscape.

Consequently, a desirable approach to address these challenges is to reduce the complexity of the model in order to improve interpretability, computational efficiency, and reliability. This can be achieved by quantifying the model's features at different scales based on practical interests and separating macroscopic scales from microscopic scales. For instance, it is always desirable to have a surrogate model that captures only the macroscopic behaviors of the full, complex model. This paper proposes an automated and systematic method for decomposing and learning effective models at an arbitrary scale specified by the user, for a given multiscale objective or energy function in multiple dimensions.

Powerful methods for scale decomposition have been developed over several decades. Many of these approaches aim to decompose the scales of a time-series signal or eliminate high-frequency component and/or noise from the time-series [15,17–19,21,25,26,29,31,48,54,66,67]. These are in some sense data-based, model-free approaches and thus are different from ours in scope.

There is also a vast body of literature dedicated to addressing multiscale problems and obtaining effective solutions or surrogate systems at specific scales of interest. Several reviews and books on multiscale modeling and simulations cover this area extensively [8,11,28,61,73]. For example, by employing appropriate numerical homogenization and discretization techniques, it is possible to

* Corresponding author.
E-mail addresses: xli47@uncc.edu (X.H. Li), mtao@gatech.edu (M. Tao).

https://doi.org/10.1016/j.jcp.2024.113206
Received 10 January 2024; Received in revised form 12 June 2024; Accepted 13 June 2024

Available online 18 June 2024
0021-9991/Published by Elsevier Inc.

directly obtain effective approximations of solutions at a chosen scale, even without explicitly obtaining the effective systems that produce these solutions. Various methods exist, including but not limited to the equation-free methods [41–43], the Heterogeneous Multiscale Methods [2,22,24,74], FLAVORs [69,70], multigrid methods [12,13,50], generalized finite element methods [7,9,27,35,68], multiscale Galerkin methods [16,23,34,36,72], operator-adapted wavelets [55,56], asymptotic-preserving methods [37,38], etc. Meanwhile, modal decomposition techniques, such as the reduced basis methods with proper orthogonal decomposition [14,33,63] and the dynamic-mode decomposition [47,64], provide effective scale decomposition of systems. These breakthroughs mainly focus on approximating the solution, as opposed to constructing approximations of the *system* that produces the solution, although we note there are also great work that construct effective surrogate models simultaneously with or after the solution approximation [7,23,33,34,47,63].

Instead of directly procuring effective solutions, numerous methods aiming at deriving surrogate models of multiscale problems have also been proposed. Here, the surrogate/effective models mainly refer to surrogate differential equations which are simpler and easier to simulate/analyze. They can, for example, be attained through averaging/homogenization [2,5,10,11,20,45,52,57,61,65]. The estimation of effective models can also be achieved via spectral information of the underlying systems [3,4,40]; or through filtering approaches and parameterized kernels [15,29,30,49,51,53,76], among others. Moreover, machine-learning approaches have become increasingly popular for constructing effective models from data. Inference learning can be conducted through statistical tools, such as maximum likelihood, methods of moments, and Bayesian inference [1,46,58], or through neural network and deep learning techniques [6,32,39,62,75].

However, many existing methods are based on analytical tools (e.g., asymptotic analysis such as averaging or homogenization) which assume an explicit scale separation. Machine learning-based methods may be less constrained by scale separation, but they typically require a good amount of training data *a priori*, making it challenging to simultaneously and flexibly learn different scales of interest.

This work proposes a novel approach that complements the existing literature, by providing a method that can decompose and learn any scale of a given multiscale objective function, in multi-dimensional Euclidean spaces. Moreover, in cases where there is only moderate or insignificant scale gaps, it can still construct some effective potential that leads to good approximations (see e.g., Sec. 4.3 for more precise statements).

The main idea of our method is leaned on an interesting fact that smaller scale components of a function can be traded for stochasticity, if one introduces an artificial step of optimizing it using gradient descent, and the scale at which this trading starts can be controlled by the learning rate of gradient descent. This idea is inspired by a recent advance in machine learning, where the problem of whether/how local minima of the training objective function can be escaped is of vital importance. Kong and Tao [44] demonstrated how gradient descent with *large* learning rate can quantitatively lead to such escapes, providing an alternative to the common escape mechanics based on noises from stochastic gradients. More precisely, the authors proved that when an objective function exhibits multiscale behaviors and is optimized by gradient descent with a large learning rate (a.k.a. time-step), the deterministic optimization dynamics acts like Langevin dynamics with potential that only describes the macroscopic part of the objective function, while its microscopic part effectively gets turned into noise via chaotic dynamics. As a result, gradient descent does not converge to a local minimizer of the objective, but instead to a statistical distribution characterized by the macroscopic component of the objective function.

Building upon this insight that a large time-step converts under-resolved scales into noise, we design a self-learning approach to learn components of a function at different scales. Firstly, we artificially introduce a damped deterministic mechanical system, specifically noiseless Langevin dynamics, with the given function as its potential. Then, we choose an appropriate time step size, corresponding to the cutoff scale above which we'd like to approximate this function, and simulate the damped mechanical system using this (large) step size. Next, we simulate the system numerically for a long time, and collect position-momentum values at different time points into a set. Points in this set will approximate a statistical distribution governed by a coarse-grained effective energy function, which however may not be Gibbs distribution yet, due to the an-isotropicity of effective noise. We will thus provide a way to estimate the covariance of the effective noise, which is nontrivial because we don't have access to the effective noise as it is only part of the underlying theory. Consequently, we will calibrate the damped mechanical system by choosing its friction matrix according to the noise covariance, and simulate the system for a long time again. This time, the collected values will approximately follow Gibbs distribution. Finally, we fit from the data the corresponding potential, which will give the effective approximation of the given function above (including) the designated scale. The resulting function can subsequently be used to construct surrogate models based on practical interests.

1.1. Problem setup and overview

Consider a function $V(\boldsymbol{q})$, which could either be the objective function of an optimization problem, as often encountered in machine learning, or in a conservation law setup, the energy function used for various physical and chemical simulations. Due to the complex nature of these application problems, $V(\boldsymbol{q})$ usually involves multiple scales. By decomposing these scales of $V(\boldsymbol{q})$, we can create a hierarchy of problems that are usually simpler to simulate, analyze, and interpret. However, it often occurs in practice that we are only given the expression of V without knowing the details of an explicit decomposition. Mathematically, we can thus assume that $V(\boldsymbol{q}) : \mathbb{R}^d \to \mathbb{R}$ with $d \geqslant 1$ *implicitly* admits a multi-scale decomposition

$$V(\boldsymbol{q}) = V_0(\boldsymbol{q}) + \sum_{j=1}^{K} V_{j,\varepsilon_j}(\boldsymbol{q}), \tag{1.1}$$

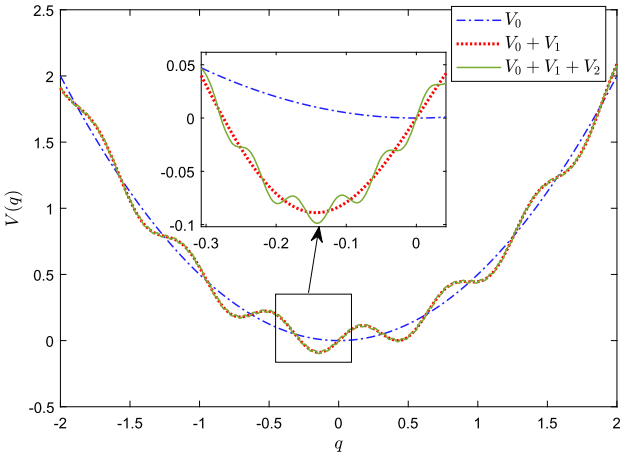


Fig. 1. The illustration of a one-dimensional potential function $V = V_0(q) + V_1(q) + V_2(q)$ with three scales: $V_0 = \frac{q^2}{2} \sim O(1)$, $V_1 = 0.1 \times \sin(\frac{q}{0.1}) \sim O(0.1)$ and $V_2 = 0.01 \times \sin(\frac{q}{0.01}) \sim O(0.01)$.

where $V_0(\cdot)$ denotes the macroscopic component, and other meso- and micro-scale components $\{V_{j,\varepsilon_j}(\cdot)\}_{j=1}^K$ satisfy $V_{j,\varepsilon_j} = O(\varepsilon_j)$ with scaling parameters $1 \gg \varepsilon_1 \gg \cdots \gg \varepsilon_K > 0$. An illustration example of $V(q)$ with three distinct scales is given in Fig. 1.

The main goals and contributions of this work are:

- Our aim is to learn each component of V at different scales, specifically V_0 and the series $\{V_{j,\varepsilon_j}\}_j$. We make the assumption that we do not possess explicit access to each individual scale, meaning the values of ε_j' are unknown *a priori*, let alone any V_{j,ε_j}.
- Once an effective potential function $U_k(\boldsymbol{q})$ that approximates V up to scale ε_k is learnt (e.g., if $\varepsilon_k \gg \varepsilon_{k+1}$, then $U_k := V_0 + \sum_{j=1}^k V_{j,\varepsilon_j}$, but note V_{j,ε_j}'s are not a priori known), various surrogate models can be constructed, and two will be focused on. In a deterministic case, a surrogate model of Hamiltonian dynamics can produce approximate phase space dynamics with less stiffness and thus allowing much larger time steps for numerical integration. In a stochastic case, a surrogate model can nearly reproduce the equilibrium distribution, auto-correlation function, and mean path of data produced by kinetic Langevin dynamics using the original $V(\boldsymbol{q})$ at the desired scale.
- In order to enable the above contributions, we have also developed an approach to estimate and calibrate kinetic Langevin systems under *an*isotropic stochastic forcings, which pose a challenge as the system's invariant distribution no longer admits an analytical expression, unlike in the isotropic case where the invariant distribution is Gibbs. This is based on estimating the covariance of noise, however without directly measuring it. Instead, we use equilibrium properties only, and by intelligently adjusting the system's mass matrix, a robust and accuracy estimation can be obtained. We can then adjust the system's friction coefficient so that the invariant distribution is put back to the analytically available Gibbs.

The organization of this paper is as follows: In Section 2, we review the theoretical foundations of proposed approach. In Section 3, we present the algorithm and explore the analytical attributes of the proposed method. The 1D case is easier as noise will always be isotropic, and it is first discussed. Then we detail the workings of the algorithm in the multivariate scenario, after describing our additional tool for kinetic Langevin estimation and calibration. Then, Section 4 delivers numerical simulations of various examples in both 1D and higher dimension. These results substantiate the robustness and efficiency of our proposed approach. Section 5 summarizes this study and suggests possible future directions.

2. Motivation and theoretical background

In this section, we will briefly review some theoretical results in [44] that motivate, and will be used by this work. We will also review some background knowledge about kinetic Langevin which will be used later.

2.1. Large step size effectively turns under-resolved scale into noise

Informal version For quantitative understanding of deep learning, Kong and Tao [44] quantified how the optimization algorithm of gradient descent escapes local minimizers of a function. They considered the optimization of a multiscale objective function by gradient descent with large learning rate, i.e. iterations given by

$$x_{k+1} = x_k - \eta \nabla f(x_k), \tag{2.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ was assumed to be admitting a macro-micro decomposition given by

$$f(x) := f_0(x) + f_{1,\epsilon}(x)$$

where f_0 is the macroscopic, $\mathcal{O}(1)$ component of the potential, and $f_{1,\epsilon}(x) = \epsilon f_1(x/\epsilon)$ is the microscopic component as it squeezes an $\mathcal{O}(1)$ function f_1 in both the x and y directions. η was referred to as the learning rate. Large learning rate is in the sense that $\eta \gg 1/L$, where L is the Lipschitz constant of ∇f, because classical analysis of gradient descent assumed $\eta < 1/L$, in which case the iterates provably converge to a (nearby) local minimizer of f.

They showed that in this case, the iteration $x_{k+1} = x_k - \eta\nabla f_0(x_k) - \eta\nabla f_1(x_k/\epsilon)$, where albeit being deterministic, exhibits behaviors similar to those of a stochastic iteration

$$x_{k+1} = x_k - \eta\nabla f(x_k) - \eta\zeta_k, \tag{2.2}$$

where ζ_k's are random variables i.i.d. to ζ which will be defined. The condition $\eta \gg 1/L$ actually leads to $\eta \gg \epsilon$, which means step size η is too large to resolve the details of $f_{1,\epsilon}$. As a consequence, the effective contribution from the microscopic, under-resolved gradients $\nabla f_1(x_k/\epsilon)$ over many iterations can be approximated by contributions from noise ζ_k. Therefore, gradient descent (2.1) with **large** learning rate η is **not** a time discretization of gradient flow ODE

$$\dot{x} = -\nabla f(x) = -\nabla f_0(x) - \nabla f_1(x/\epsilon)$$

but can be intuitively understood as a stepsize η time discretization of **SDE**

$$dx = -\nabla f_0(x)dt + \sqrt{\eta}\Sigma dW_t \tag{2.3}$$

for some constant covariance matrix Σ given by $\Sigma\Sigma^T = \mathbb{E}[\zeta\zeta^T]$.

Selected rigorous details They considered $\epsilon \ll 1$, f_0, $f_{1,\epsilon} \in C^2(\mathbb{R}^d)$ and microscopic potentials $f_{1,\epsilon}$ satisfying the following two conditions:

- $\mathcal{O}(1)$ **gradient condition:** there exists a bounded and nonconstant random variable (r.v.) $\zeta \in \mathbb{R}^d$ with $\mathbb{E}(\zeta) = 0$, such that: $\forall\epsilon > 0$ and $\forall x \in \mathbb{R}^d$ there exists a positive measured set $\Gamma_{x,\epsilon} \subset B(x, \delta(\epsilon)$ with $\lim_{\epsilon\downarrow 0}\delta(\epsilon) = 0$, such that a uniformly distributed r.v. on $\Gamma_{x,\epsilon}$, denoted by $u_{\Gamma_{x,\epsilon}}$, satisfies $\nabla f_{1,\epsilon}(x + u_{\Gamma_{x,\epsilon}}) \xrightarrow{w} -\zeta$ as $\epsilon \to 0$, uniformly with respect to x.
- $\mathcal{O}(1/\epsilon)$ **Hessian condition:** $\epsilon\nabla^2 f_{1,\epsilon}$ is uniformly bounded as $\epsilon \to 0$. Also $\exists m > 0$, such that for any bounded and nonempty rectangle $\Gamma \subset \mathbb{R}^d$, the expectation $\mathbb{E}\left[\ln\|\epsilon\nabla^2 f_{1,\epsilon}(u_\Gamma)\|_2\right] \to m$ as $\epsilon \to 0$, where u_Γ is a uniform r.v. on Γ.

While these conditions are nontrivial, Kong and Tao noted that they include but are strictly weaker than periodicity (e.g., $f_{1,\epsilon}(x) = \epsilon\sin(x/\epsilon)$, for which $\Sigma = 1/\sqrt{2}$) or quasiperiodicity.

They then considered the deterministic dynamical system induced by iteration map

$$\varphi : x \mapsto x - \eta\nabla f(x) = x - \eta\nabla_0 f(x) - \eta\nabla f_{1,\epsilon}(x), \tag{2.4}$$

and proved the following results (selected):

1. Fix η and let $\epsilon \to 0$. If φ has a family of nondegenerate invariant distributions for $\{\epsilon_i\}_{i=1}^{\infty} \to 0$ which converges in the weak sense, then the weak limit is an invariant distribution of $\hat{\varphi}$, where $\hat{\varphi}$ defines a stochastic map

$$\hat{\varphi} : x \mapsto x - \eta\nabla f_0(x) + \eta\zeta. \tag{2.5}$$

2. In a special case where the macroscopic potential f_0 is strongly convex and L-smooth, there exists some constant $C > 0$ independent of ϵ, such that the mapping $\hat{\varphi}$ has a unique invariant distribution for any fixed $\eta \leqslant C$ and the iteration of $\hat{\varphi}$ converges exponentially fast to this distribution. Moreover, if the covariance matrix of ζ is isotropic, i.e., $\sigma^2 \mathrm{I}_d$, then the rescaled Gibbs distribution $\frac{1}{Z}\exp\left(-\frac{2f_0(x)}{\eta\sigma^2}\right)\mathrm{d}x$ is an $O(\eta^2)$ approximation of that of $\hat{\varphi}$.

However, it is worth noting that [44] only considered two scales in the objective function $f(x)$, whereas this work heuristically extends the application to arbitrary K scales. In addition, note ζ_k in (2.2) is actually bounded, unlike 'dW_t' in (2.3), but (2.3) is still a reasonable interpretation in a central limit theorem sense (see [44] for details). Additionally, the stochastic behavior of the deterministic mapping with a large LR was only proved for gradient descent dynamics, and if one considers a damped deterministic mechanical system instead, with ∇f being its forcing, and discretize its time using a stepsize that under-resolves part of f, the quantitative stochasticity of its dynamics was only a conjecture. While this work relies on this conjecture, our empirical results appear to be consistent with it, namely deterministic simulations of damped mechanical systems with large step size exhibit behaviors resemblant to (stochastic) kinetic Langevin dynamics.

2.2. Review of kinetic Langevin dynamics

Kinetic Langevin equation models how particles evolve subject to a combination of conservative forcing, damping, and thermal noise. Let $(\boldsymbol{q}, \boldsymbol{p})$ denote the position and momentum of particles, and M denote the corresponding mass matrix, then it gives the evolution of $(\boldsymbol{q}, \boldsymbol{p})$ by

$$\begin{cases} d\boldsymbol{q} = M^{-1}\boldsymbol{p}dt, \\ d\boldsymbol{p} = (-\nabla V(\boldsymbol{q}) - \Gamma M^{-1}\boldsymbol{p})dt + \Sigma d\boldsymbol{W}_t, \end{cases} \tag{2.6}$$

where ∇V denotes the potential force, $\boldsymbol{W}_t$ denotes the standard multivariate Wiener process, M is a positive definite mass matrix, Γ is the friction coefficient matrix for the dissipative force, and Σ denotes the diffusion coefficient matrix for the thermal random forces.

It is known (e.g., [59]) that if the diffusion coefficient matrix Σ and the friction coefficient matrix Γ satisfy the following fluctuation–dissipation relation

$$\Sigma\Sigma^T = 2\Gamma, \tag{2.7}$$

then the equilibrium distribution is a Gibbs distribution

$$d\pi(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{Z}\exp -\left(\left(V(\boldsymbol{q}) + \boldsymbol{p}^T M^{-1}\boldsymbol{p}/2\right)\right) d\boldsymbol{q}d\boldsymbol{p}. \tag{2.8}$$

The most commonly discussed case is when the thermal noises are isotropic, that is $\Sigma\Sigma^T = \sigma^2 \mathrm{I}_d$ for some scalar σ. In this case, the equilibrium distribution is always Gibbs as long as $\Gamma = \frac{\sigma^2}{2}\mathrm{I}_d = \gamma \mathrm{I}_d$.

However, if the noises are anistropic, the invariant distribution of (2.6) still exists under mild conditions, but it may not admit an analytical expression.

3. Main results

In this section, we propose a self-learning algorithm that automatically identifies macroscopic scales of a function V (eq. (1.1)) up to a desired scale ε. Here, ε corresponds to a cutoff level k with $1 \leqslant k \leqslant K$ satisfying $\varepsilon_k \gtrsim \varepsilon \gg \varepsilon_{k+1}$, meaning we seek $U_k = V_0 + \sum_{j=1}^{k} V_{j,\varepsilon_j}$. Of course, if the goal is to separate all the scales instead, we can just repeat this procedure for different ε values.

3.1. The proposed self-learning methodology for multi-dimensions

Our first step of identifying components of V above scale ε, solely based on V,[1] is to extend the gradient descent iterations analyzed in [44] (see Sec. 2.1 for a quick summary) by including momentum.

More precisely, we consider a large step size simulation of the following damped mechanical system

$$\begin{cases} \dot{\boldsymbol{q}} = M^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} = -\nabla V(\boldsymbol{q}) - \Gamma M^{-1}\boldsymbol{p}, \end{cases} \tag{3.1}$$

where the step size is chosen to be at the order of ε, i.e. $\delta \sim \varepsilon$, so that it underresolves undesired smaller scales. Here M and Γ are mass and friction coefficient matrices that we can choose, but $V(\boldsymbol{q})$ is given and we have to use it in its entirety as its scale decomposition is not yet known (this actually poses a significant challenge in high dimension in general, which will be described and addressed in Sec. 3.2).

The choice of numerical integrator for (3.1) for the purpose of turning unresolved small scales effectively into 'noise' needs not to be unique, but here we simply apply a dissipative generalization of Störmer–Verlet, where friction is handled by an exponential integrator. The evolution of discrete solution $(\boldsymbol{q}_n, \boldsymbol{p}_n)$ at time grid n thus becomes

$$\begin{aligned} \boldsymbol{q}_{n+\frac{1}{2}} &= \boldsymbol{q}_n + M^{-1}\frac{\delta}{2}\boldsymbol{p}_n, \\ \boldsymbol{p}_{n+1} &= e^{-\Gamma M^{-1}\delta}\boldsymbol{p}_n - \delta\nabla V(\boldsymbol{q}_{n+\frac{1}{2}}), \\ \boldsymbol{q}_{n+1} &= \boldsymbol{q}_{n+\frac{1}{2}} + M^{-1}\frac{\delta}{2}\boldsymbol{p}_{n+1}, \end{aligned} \tag{3.2}$$

and with given initial condition $(\boldsymbol{q}_0, \boldsymbol{p}_0)$ from a suitable region. We conjecture that the gradient of under-resolved microscopic components of V can effectively be approximated by noises as $\varepsilon_{k+1} \to 0$, similar to Sec. 2.1. This conjecture means that invariant distribution of the deterministic dynamics can be approximated by that of the stochastic system

[1] More precisely, only 1st-order oracle, no additional data or information.

$$\begin{cases} d\boldsymbol{q} = M^{-1}\boldsymbol{p}dt, \\ d\boldsymbol{p} = (-\nabla U_k(\boldsymbol{q}) - \Gamma M^{-1}\boldsymbol{p})dt + \Sigma dW_t, \end{cases} \tag{3.3}$$

with the effective potential $U_k = V_0 + \sum_{j=1}^{k} V_{j,\varepsilon_j}$, $\Gamma = \frac{1}{2}\Sigma\Sigma^T$, and the diffusion coefficient matrix given by

$$\Sigma\Sigma^T := \delta\mathbb{E}\left[\nabla\left(\sum_{j=k+1}^{K} V_{j,\varepsilon_j}(\boldsymbol{u})\right)\otimes\nabla\left(\sum_{j=k+1}^{K} V_{j,\varepsilon_j}(\boldsymbol{u})\right)\right], \tag{3.4}$$

where the expectation is with respect to an auxiliary random variable $\boldsymbol{u}$ defined in the gradient conditions in Section 2, i.e. a uniformly distributed r.v. on any bounded rectangle in $\mathbb{R}^d$ with size independent of $\varepsilon_{k+1}, \cdots, \varepsilon_K$, i.e. $\mathcal{O}(1)$. The sense of approximation is that the invariant distribution of (3.3) differs from that of (3.2) by at most $\mathcal{O}(\delta)$ (in weak* topology) as scale separation goes to infinity (i.e. $\delta/\varepsilon_{k+1} \to \infty$).

In addition, if the system is mixing, (unfortunately, the precise necessary and sufficient condition for so is still a major open problem, but it was conjectured to be the case for large learning rate gradient descent; see Section 2 for semi-quantitative discussions), then (3.2) will converge to its invariant/equilibrium distribution, denoted by π_δ. This in the sense that 1) an ensemble of trajectories with random initial condition following any smooth density will converge to the invariant distribution as $n \to \infty$, and 2) except for measure zero initial conditions, any single trajectory will have an ergodic limit with respect to the invariant distribution π_δ, meaning for any smooth test function ϕ, we have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \phi(\boldsymbol{q}_n, \boldsymbol{p}_n) = \mathbb{E}_{\boldsymbol{q},\boldsymbol{p}\sim\pi_\delta}\phi(\boldsymbol{q},\boldsymbol{p}).$$

This means, given data $\{\boldsymbol{q}_n, \boldsymbol{p}_n\}$ generated from V via (3.2), the empirical distribution of $\{\boldsymbol{q}_n, \boldsymbol{p}_n\}_{n=1,\cdots,N}$ converges to π_δ, which is an $\mathcal{O}(\delta)$ approximation of π, whose density is $Z^{-1}\exp(-U_k(\boldsymbol{q}) - \boldsymbol{p}^T M^{-1}\boldsymbol{p}/2)$ for (3.3).

Therefore, for a given Γ, Step Two of our approach is to run the deterministic system (3.2) for sufficiently long on time interval $[0, T]$ until $(\boldsymbol{q}_n, \boldsymbol{p}_n)$ reach their statistical equilibrium, which can be tested from the normality test [71] of the distribution of $\{\boldsymbol{p}_n\}$. Note our introduction of momentum is rather important, despite the fact that the whole dynamics is just artificially introduced, because the convergence of $\boldsymbol{p}$ to simple Gaussian enables accurate detection of near convergence.

This Step Two generates a training data set solely out of V, which will be used later for the learning of an effective macroscopic function, hence the name 'self-learning'.

Remark 3.1. Due to the addition of momentum considered in this work, we could not establish conditions like those in [44] on the gradient and Hessian and prove convergence to near Gibbs under these conditions, because such a proof seems to require a completely different machinery. Even if we could replace the bounded, correlated, and anisotropic microscopic gradients in our dynamics (3.1) and (3.2) by unbounded, uncorrelated, and isotropic white noise process, which would turn the dynamics into kinetic Langevin, the convergence remains more challenging to prove compared to the overdamped Langevin. Meanwhile, our results seem very robust in all sorts of numerical experiments in Section 4.

Remark 3.2. Although in principle it is very flexible to choose the initial condition $(\boldsymbol{q}_0, \boldsymbol{p}_0)$ because of the stochasticity induced by microsuitable large step size, it becomes subtle in practice as good initial conditions can greatly improve the exploration of the dynamics and speed-up the convergence of the algorithm. We will provide more details for specific examples in Section 4.

Remark 3.3. In Step Two, the normality test is suggested on the distribution of $\{\boldsymbol{p}_n\}$ to check if the system reach its statistical equilibrium. However, it is only a necessary condition but not always sufficient. This test is more trustworthy when the temperature parameter $1/(\hat{\beta})$ is comparable to or greater than the energy barriers of the objective function $V(\boldsymbol{q})$. For systems with higher energy barriers, global exploration is usually lost. Lowering $\hat{\beta}$ too much introduces statistical noise that overwhelms the information about $\{U_k\}$. Otherwise, the simulation may lead to the system relaxing into one metastable state (with $\boldsymbol{p}$ following a normal distribution), but $\boldsymbol{q}$ never explores other metastable states. Therefore, we only consider the objective function V_0 with energy barriers of order $O(1)$.

Remark 3.4. We explain the strategy here when the energy barriers of V_0 are of $O(1)$ and the scale separation among $\{\varepsilon_j\}_{j=1}^K$ exists with range of scales knowing, but the value of each ε_j keeps unknown.

The idea is to firstly use an isotropic friction matrix $\Gamma = \gamma I$ and run the damped mechanical system (3.2) using δ from an array of time-stepping sizes, which are arranged in a decreasing order and are compatible with the range of scales of $V(\boldsymbol{q})$. For each selected δ, we simulate long enough until numerically reach the equilibrium distribution of $\pi_\delta(\boldsymbol{q}, \boldsymbol{p})$. If the distribution of $\boldsymbol{p}$ becomes a general multi-dimensional normal distribution via a normality test, then it suggests we find a scale ε_k which is proportional to δ. Otherwise, if the distribution of $\boldsymbol{p}$ is far away from a normal distribution, then this suggests the system is still under mixed scales and we will further decease δ and regenerate data. After we find the proper δ at a targeting scale, we next tune the friction matrix Γ to achieve the Gibbs distribution (3.6).

The successful results of a benchmark test with unknown scales can be found in Fig. 7. Also, notice that the simulations on different δ are parallelizable, so the total efficiency of distinguish scales will be ensured.

3.2. Microscopic covariance estimation and the creation of fluctuation-dissipation balance

Recall that the friction coefficient Γ is a parameter we can choose, while Σ is something we cannot control because it characters the effective noise that originates from the microscopic gradient. If we can align Γ with Σ so that they commute, or more precisely so that $\Sigma\Sigma^T = 2\Gamma/\hat{\beta}$ for some constant scalar $\hat{\beta}$ (known as the inverse temperature), then the data collected in the above Step 2 will approximately follow a density $\propto \exp\left(-\hat{\beta}\left(U_k(\boldsymbol{q}) + \boldsymbol{p}^T M^{-1}\boldsymbol{p}/2\right)\right)$, which means we can recover U_k via regression of the $\boldsymbol{q}$ marginal of the data.

However, we do not know Σ a priori because we only have V, but not the microscopic potential U_k, upon which Σ is based. This thus poses a challenge.

When V is one-dimensional, this challenge is relatively easy to solve, as aligning Γ with Σ is not an issue and we just need the inverse temperature, which can be estimated from $\mathbf{p}$'s variance. When V is a multi-dimensional function in general, choosing Γ so that the system admits Gibbs as its invariant distribution is rather nontrivial, but we will provide a rigorous and systematic method. Both cases are now detailed:

One-dimensional strategy When $d = 1$, the friction matrix Γ becomes just a scalar γ. Hence, the equilibrium distribution of (q, p) is always a Gibbs distribution. Consequently, M and γ can be selected easily in (3.1). Throughout the paper, we select the following values for the 1D simulation $0 < \gamma \leqslant 0.5$ and $M = 1$. Once the numerical solution $\{(q_n, p_n)\}_n$ is nearly in the statistical equilibrium, the distribution of p_n will be close to a Gaussian distribution. Following [60], we estimate the equivalent thermal noise effect $\hat{\beta}$ induced by the gradient of micro-scale components

$$\hat{\beta} = \frac{1}{\mathsf{var}(p)}. \tag{3.5}$$

In 1D, $\pi_\delta(q, p)$ is approximated by

$$\pi_\delta(q, p) \sim \exp\left(-\hat{\beta}\big(U_k(q) + \frac{p^2}{2}\big)\right) dq\, dp. \tag{3.6}$$

In 1D simulation with suitable value of γ, we can get $\hat{\beta}$ in the range around $\hat{\beta} \in (0.2, 1.25)$, which will lead to better learning results for U_k.

Multi-dimensional strategy As mentioned above, due to the anisotropy of microscopic scales in $V(\boldsymbol{q})$ the dimension $d \geqslant 2$, invariant distribution of kinetic Langevin (3.3) in general may not be interpretable. Therefore, we construct a special case by estimating Σ (originated from the microscopic scales (3.4)) and choose $\Gamma = \frac{1}{2}\Sigma\Sigma^T$ to ensure the fluctuation-dissipation relation (2.7). In this case, the temperature is always $\hat{\beta} = 1$. We explain the details by a two-stage scheme.

Stage 1 and 2: apply various M to estimate $\Sigma\Sigma^T$ The key idea of our innovation is to exploit the effects of different choices of mass matrix M in the following stochastic Langevin dynamics:

$$\begin{cases} d\boldsymbol{q} = M^{-1}\boldsymbol{p}dt, \\ d\boldsymbol{p} = \left(-\gamma M^{-1}\boldsymbol{p} - \nabla V(\boldsymbol{q})\right) dt + \Sigma d\boldsymbol{W}_t, \end{cases} \tag{3.7}$$

where $0 < \gamma$ is some scalar friction constant. Note $\nabla V\, dt + \Sigma d\boldsymbol{W}_t$ come together from the under-resolved small scales, and thus again Σ cannot be chosen or determined a priori. The positive-definite mass matrix M however can be arbitrarily chosen, because (3.7) is just some auxiliary dynamics designed for obtaining an algorithm. We will design M values so that Σ can be estimated from 'solutions' of (3.7), which are actually collected as the deterministic iterations with step size δ

$$\begin{aligned} \boldsymbol{q}_{n+\frac{1}{2}} &= \boldsymbol{q}_n + M^{-1}\frac{\delta}{2}\boldsymbol{p}_n, \\ \boldsymbol{p}_{n+1} &= e^{-\gamma M^{-1}\delta}\boldsymbol{p}_n - \delta\nabla V(\boldsymbol{q}_{n+\frac{1}{2}}), \\ \boldsymbol{q}_{n+1} &= \boldsymbol{q}_{n+\frac{1}{2}} + M^{-1}\frac{\delta}{2}\boldsymbol{p}_{n+1}. \end{aligned}$$

Meanwhile, the relation between $\Sigma\Sigma^T$ and the solutions is established in the following proposition.

Theorem 3.1. *Consider a stochastic Langevin dynamics (3.7) with a positive scalar friction constant γ and a constant diffusion coefficient matrix Σ, then Σ satisfies*

$$\mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) = \gamma\mathbb{E}\left[\boldsymbol{p}^T M^{-2}\boldsymbol{p}\right], \tag{3.8}$$

where the expectation is taken with respect to the invariant distribution of $(\boldsymbol{q}, \boldsymbol{p})$.

Proof. We recall the definition of Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} + V(\boldsymbol{q})$ and evaluate it along the solutions of (3.7). Itô formula shows that the Hamiltonian satisfies the following SDE

$$\begin{aligned} dH &= (M^{-1}\boldsymbol{p})^T \left[-\gamma M^{-1}\boldsymbol{p} - \nabla V(q)\right] dt + \nabla V(q)^T M^{-1}\boldsymbol{p} dt + \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) dt + \left(M^{-1}\boldsymbol{p}\right)^T \sqrt{2}\Sigma d\boldsymbol{W}_t \\ &= -\gamma \boldsymbol{p}^T M^{-2}\boldsymbol{p} dt + \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) dt + \sqrt{2}\boldsymbol{p}^T M^{-1}\Sigma d\boldsymbol{W}_t. \end{aligned}$$

Integrating it from time 0 to t, we get

$$H(t) = H(0) + \int_0^t -\gamma \boldsymbol{p}^T(\tau)M^{-2}\boldsymbol{p}(\tau)d\tau + \int_0^t \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) d\tau + \sqrt{2}\int_0^t \boldsymbol{p}^T(\tau)M^{-1}\Sigma d\boldsymbol{W}_\tau.$$

We take expectation of H with respect to the invariant distribution and notice that the last Itô integral term is a martingale, so

$$\mathbb{E}(H(t)) =: E(t) = E(0) + \int_0^t -\gamma\mathbb{E}\left[\boldsymbol{p}^T(\tau)M^{-2}\boldsymbol{p}(\tau)\right] d\tau + \int_0^t \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) d\tau.$$

Hence we get the differential equation of $\mathbb{E}(H(t))$, which is

$$\dot{E}(t) = -\gamma\mathbb{E}\left[\boldsymbol{p}^T(t)M^{-2}\boldsymbol{p}(t)\right] + \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right).$$

When the system reaches its statistical equilibrium, we have $\dot{E}(t) = 0$, so

$$0 = \dot{E}(t) = -\gamma\mathbb{E}\left[\boldsymbol{p}^T(t)M^{-2}\boldsymbol{p}(t)\right] + \mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right).$$

Consequently, we can set up a relation between M and Σ at the statistical equilibrium of system, that is

$$\mathsf{Tr}\left((\Sigma\Sigma^T)M^{-1}\right) = \gamma\mathbb{E}\left[\boldsymbol{p}^T M^{-2}\boldsymbol{p}\right],$$

which proved the proposition. □

In order to estimate $Z := \left(\Sigma\Sigma^T\right) = \left(z_{ij}\right)_{d\times d}$, we then can design the entries of inverse of mass matrix $A := M^{-1} = \left(a_{ij}\right)$ accordingly and apply the following steps:

- Step 1: choose a set of d different diagonal matrices $A^{(k)} = \mathsf{diag}(a_{ii}^{(k)})$ and run the Langevin dynamics (3.7) to reach its statistical equilibrium. Thus, we get a system of d linear equations to solve all diagonal entries z_{ii}:

$$\sum_{i=1}^{d} a_{ii}^{(k)} z_{ii} = \gamma\mathbb{E}\left[\boldsymbol{p}^{(k)^T} M^{-2}\boldsymbol{p}^{(k)}\right], \quad k = 1, \dots, d.$$

- Step 2: fix a pair of off-diagonal entries with $\ell \neq r$, set this pair $a_{r\ell} = a_{\ell r} = 1/2$, set all diagonal entries $a_{ii} = 1$, and set the rest of off-diagonal entries $a_{ij} = 0$, then run the Langevin dynamics (3.7) until reaching the equilibrium. We thus have

$$\left(z_{\ell r} + \sum_{i=1}^{d} z_{ii}\right) = \gamma\mathbb{E}\left[\boldsymbol{p}^T M^{-2}\boldsymbol{p}\right],$$

  which can solve $z_{\ell r}$. Hence, after repeating this procedure for $\frac{d(d-1)}{2}$ times, we can get all values of the off-diagonal entries of $Z = (\Sigma\Sigma^T)$.

Combining step 1 with step 2, overall we need to solve $\frac{d(d+1)}{2}$ linear systems. However, each one of them is independent of the others which means all the simulations can be run in parallel.

Remark 3.5. To demonstrate the scheme in details, we consider a two-dimensional case $d = 2$.

Stage 1: apply various M. In step 1, we choose the two inverse of mass matrices to be

$$A^{(1)} = M^{-1,(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad A^{(2)} = M^{-1,(2)} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

In step 2, we choose the inverse of mass matrices to be

$$A^{(3)} = M^{-1,(3)} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

Stage 2: estimate Γ. Once we get the estimation of Σ from stage 1, we set the mass matrix to be $M = \mathbb{I}_d$ and set the friction coefficient matrix Γ to be

$$2\Gamma = \Sigma\Sigma^T = Z.$$

Input: Target function $V(\boldsymbol{q})$, a step size $\delta \sim O(\varepsilon_k)$.
Output: Effective function $U_k(\boldsymbol{q}) \sim O(\varepsilon_k)$ and effective covariance matrix $\Sigma\Sigma^T$ from unresolved scales.
1: **Stage 1 and 2 :** estimate $\Sigma\Sigma^T$ and Γ:
Set $\Gamma = \gamma I$ with $0 < \gamma \leqslant 0.1$, fix the step size δ, run (3.2) with $V(\boldsymbol{q})$ for a total of $\frac{d(d+1)}{2}$ independent trajectories using different mass matrices. Solve entries of $\Sigma\Sigma^T$ via (3.8). Set $\Gamma = \Sigma\Sigma^T/2$.
2: Estimate $\pi_\delta(\boldsymbol{q})$ and $U_k(\boldsymbol{q})$:
Set $M = I$, fix the step size δ, run (3.2) with $V(\boldsymbol{q})$ for one trajectory. Estimate the empirical distribution $\pi_\delta(\boldsymbol{q})$ from the data $\{\boldsymbol{q}_n\}$. Learn $U_k(\boldsymbol{q})$ via (3.6).

Algorithm 1. Learn effective function $V = U_k + \mathcal{O}(\varepsilon_k)$.

3.3. Estimate the effective potential

We apply the Verlet scheme (3.2) to the damped mechanical system (3.1) (not kinetic Langevin), with appropriate Γ, for sufficiently long, and record the trajectory as $\{\boldsymbol{q}_n, \boldsymbol{p}_n\}$, i.e. our data. We propose to learn the effective potential U_k from the invariant distribution (3.6), which is

$$\log\left(\pi_\delta(\boldsymbol{q})\right) = -\log Z + \hat{\beta} U_k(\boldsymbol{q}), \tag{3.9}$$

where $\pi_\delta(\boldsymbol{q})$ is the marginal of invariant distribution, approximated empirically by $\{\boldsymbol{q}_n\}$. For 1D case, $\hat{\beta} = \frac{1}{\mathsf{var}(p)}$, while for multi-dimensional case $\hat{\beta} = 1$ as Γ is set to satisfy $\Gamma = \frac{1}{2}\Sigma\Sigma^T$.

Learning $U_k(\boldsymbol{q})$ from data via (3.9) is a function approximation / interpolation problem. There are of course many ways to interpolate the data, but since function approximation is not the main point of this paper, we will just pick one approach. Specifically, we choose a set of suitable basis functions, for instance the piecewise spline basis, $\{\varphi_i(\boldsymbol{q})\}_i^m$ to approximate $U_k(\boldsymbol{q})$ by

$$U_k(\boldsymbol{q}) \approx \sum_{i=1}^{m} a_i \varphi_i(\boldsymbol{q}),$$

where we can solve a regression problem to learn the coefficients $\{a_i\}_{i=1}^m$.

Remark 3.6. To apply the method to high-dimensional scenarios, there are at least two aspects that should be considered. One is to estimate a probability distribution from data, and the other is the computational cost of covariance matrix estimation. The dimension dependence for covariance matrix estimation should be polynomial, and this part is thus not a curse of dimensionality because it is not an exponential dependence. The density estimation, on the other hand, is an important, well-known statistical problem. Whether there is a curse of dimensionality will depend on the distribution and the estimation method. We feel these questions, albeit important, are beyond the scope of this work, and in this paper we just use some naive method (polynomial fitting), which does face a curse of dimensionality.

3.4. Summary of the algorithm

See Algorithm 1.

4. Surrogate models and numerical experiments

We now numerically test the efficacy of our approach in two senses. One is about the accuracy of the learned effective function $U_k(\boldsymbol{q})$. The other is about the approximation abilities of derived surrogate models; i.e., if the learned U_k replaces the full V when used in a dynamical setup, how would the resulting dynamics differ. For this latter point, we will illustrate two dynamical setups, one being Hamiltonian dynamics and the other being kinetic Langevin, in both cases V or U_k will be the potential. The versions with U_k are the surrogate models, whose benefits are that they can be simulated using larger setups, easier to analyze, and sometimes more intuitive to interpret as well.

More precisely, the surrogate models are given by

$$\textbf{Hamiltonian:} \begin{cases} \dot{\boldsymbol{q}} = M^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} = -\nabla U_k(\boldsymbol{q}), \end{cases} \tag{4.1}$$

and

$$\textbf{Kinetic Langevin:} \begin{cases} \dot{\boldsymbol{q}} = M^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} = -\nabla U_k(\boldsymbol{q}) - \Gamma M^{-1}\boldsymbol{p} + \sqrt{2\Gamma}\, d\boldsymbol{W}_t. \end{cases} \tag{4.2}$$

For the comparison of the deterministic Hamiltonian system, we will mainly focus on the comparison of the phase portraits $(\boldsymbol{q}, \boldsymbol{p})$ generated by $V(\boldsymbol{q})$ and $U_k(\boldsymbol{q})$. For the case of kinetic Langevin, we will compare the equilibrium distribution, mean path (ensemble average of trajectories) and normalized auto-correlation functions (Normalized ACF) of $\boldsymbol{q}$. Note that this is not a tautology, as we will not only inspect the equilibrium aspect of the surrogate model, which should match that of the full model as long as the proposed

trading of small-scale for effective noise is correct, but we will also investigate out-of-equilibrium dynamical aspects by comparing the mean trajectories and normalized ACF.

(Empirical) mean trajectory is defined as

$$\bar{\boldsymbol{q}}(t_n) = \frac{1}{M}\sum_{\ell=1}^{M} \boldsymbol{q}^{(\ell)}(t_n), \tag{4.3}$$

for which we evolve M independent trajectories from i.i.d. random initial condition, and $\boldsymbol{q}^{(\ell)}(t_n)$ denotes the ℓ-th one of them. We also define the normalized ACF as

$$\text{Normalized ACF}(\tau) = \frac{1}{\widetilde{\text{Cov}}(\boldsymbol{q})}\left(\widetilde{\mathbb{E}}[\boldsymbol{q}\,\boldsymbol{q}_{\cdot+\tau}] - \widetilde{\mathbb{E}}[\boldsymbol{q}]\,\widetilde{\mathbb{E}}[\boldsymbol{q}_{\cdot+\tau}]\right), \tag{4.4}$$

where $\widetilde{\text{Cov}}(\boldsymbol{q})$ denotes the covariance of $\boldsymbol{q}$ across time as well as the ensembles; and $\widetilde{\mathbb{E}}$ denotes the time average with respect to the ensembles. Here, $\widetilde{\mathbb{E}}[\boldsymbol{q}]$ approximates $\widehat{\mathbb{E}}[\boldsymbol{q}]$, which in the continuous case is defined as

$$\widehat{\mathbb{E}}[\boldsymbol{q}] := \lim_{T\to\infty}\frac{1}{M}\sum_{\ell=1}^{M}\left(\frac{1}{T}\int_0^T \boldsymbol{q}_t^{(\ell)}\,dt\right) \quad \text{and} \quad \widehat{\mathbb{E}}[\boldsymbol{q}_{\cdot+\tau}] := \lim_{T\to\infty}\frac{1}{M}\sum_{\ell=1}^{M}\left(\frac{1}{T}\int_0^T \boldsymbol{q}_{t+\tau}^{(\ell)}\,dt\right).$$

In the discrete case, an empirical approximation for uniform time-stepping size is defined as

$$\widetilde{\mathbb{E}}[\boldsymbol{q}] := \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{q}_{t_n}^{(\ell)} \quad \text{and} \quad \widetilde{\mathbb{E}}[\boldsymbol{q}_{\cdot+\tau}] := \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{q}_{t_n+n_\tau}^{(\ell)},$$

with n_τ being fixed and satisfying $\tau = \delta n_\tau$.

Based on the definition of $\widetilde{E}$, both $\widetilde{\mathbb{E}}[\boldsymbol{q}\,\boldsymbol{q}_{\cdot+\tau}]$ and $\widetilde{\text{Cov}}$ are also computed via the empirical approximation across time and ensembles:

$$\widetilde{\mathbb{E}}[\boldsymbol{q}\,\boldsymbol{q}_{\cdot+\tau}] := \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{q}_{t_n}^{(\ell)}\boldsymbol{q}_{t_n+n_\tau}^{(\ell)}, \quad \text{and} \quad \widetilde{\text{Cov}}(\boldsymbol{q}) := \widetilde{\mathbb{E}}\left[\left(\boldsymbol{q}-\widetilde{\mathbb{E}}[\boldsymbol{q}]\right)\left(\boldsymbol{q}-\widetilde{\mathbb{E}}[\boldsymbol{q}]\right)\right].$$

In the following examples, we will consider two cases: 1) values of $\{\varepsilon_k\}$ are explicitly known, and 2) values of $\{\varepsilon_k\}$ are not explicitly known.

4.1. Parameters for numerical tests

Here, we list some numerical parameters which are commonly used in the following numerical tests. For most 1D examples, we fix the total number of time steps to be $N_t = 5\times 10^9$ and we set the scalar friction constant to be $\gamma = 0.1$. For 2D examples, we fix the total number of time steps to be $N_t = 2\times 10^7$, and use the Algorithm 1 to match the friction matrix and diffusion coefficient according to (2.7). In all examples of both dimensions, the time step size δ will be chosen according to the scale of interest ε_k. When considering the effectiveness of surrogate Hamiltonian (4.1), we only generate one trajectory to compare the results of phase portrait. On the other hand, when comparing the Langevin system (4.2), we will use standard normal distributed initial conditions for $(\boldsymbol{q}, \boldsymbol{p})$ to generate $M = 4000$ independent trajectories on time interval $[0, T]$.

4.2. One-dimensional examples with values of scales known

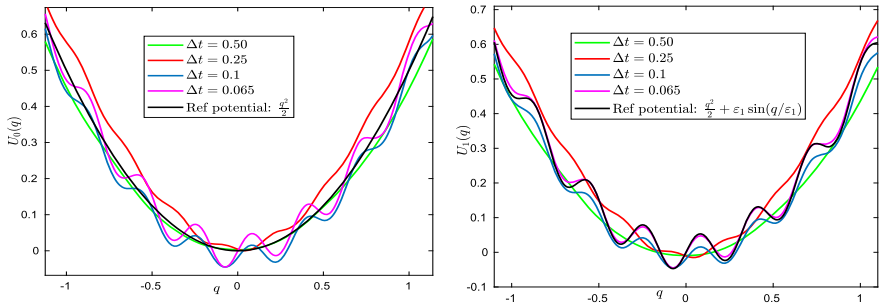
1. **Test 1: quadratic potential.** We first consider a simple three-scale function, given by a quadratic $\mathcal{O}(1)$ component modulated by two smaller periodic scales

$$V(q) = \frac{q^2}{2} + \varepsilon_1\sin(q/\varepsilon_1) + \varepsilon_2\sin(q/\varepsilon_2) \text{ with } (\varepsilon_1, \varepsilon_2) = (0.05, 0.001). \tag{4.5}$$

We run the scheme (3.2) with various time step sizes δ and plot the learnt functions in Fig. 2. Notice that for $\delta = 0.5$, the learned potential greatly resembles $U_0 = \frac{q^2}{2}$, and for $\delta = 0.065$, the learned potential greatly resembles $U_1 = \frac{q^2}{2} + 0.05\sin(q/0.05)$. Clearly, numerical step size as δ acts as a scaling filter to select the resolution of observed data.

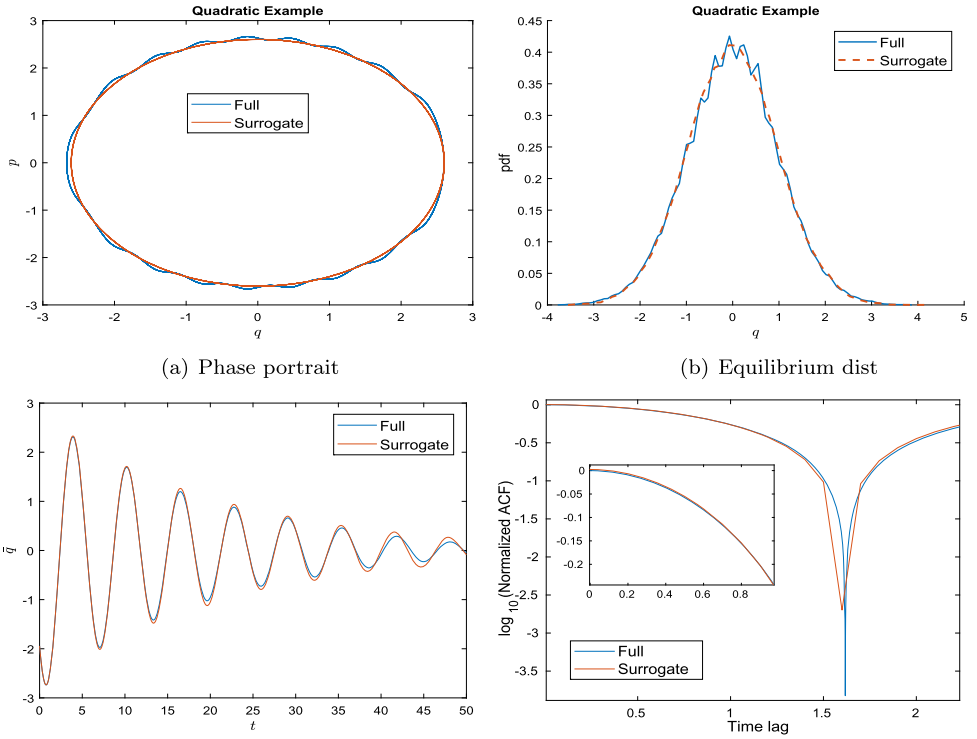
Based on learnt U_0, we run the surrogate Hamiltonian (4.1) and the surrogate Langevin system (4.2) using $U_0(q)$ with $\delta = 0.1$, and compare the simulations results using $V(q)$ with $h = 5e-4$ for the Hamiltonian and Langevin dynamics, respectively. As mentioned above, we compare the phase portrait for the Hamiltonian system, the equilibrium distribution, mean trajectory (4.3) and normalized ACF (4.4). The results are summarized in Fig. 3. For this quadratic example, we see that the surrogate models using $U_0(k)$ work well regarding capturing both statistics.

2. **Test 2: double-well potential.** We next consider a three-scale function with its macroscopic component being a non-convex double-well function, instead of the previously considered convex quadratic function, and assume that we know the two small scales explicitly



(a) Reference $\frac{q^2}{2}$ (b) Reference $\frac{q^2}{2} + \varepsilon_1 \sin(q/\varepsilon_1)$

Fig. 2. The plots of learnt functions from simulated data with various time step sizes δ and the reference function $U_k(q)$. The exact function of full scales is $V(q) = \frac{q^2}{2} + \varepsilon_1 \sin(q/\varepsilon_1) + \varepsilon_2 \sin(q/\varepsilon_2)$. Parameters are $(\varepsilon_1, \varepsilon_2) = (0.05, 0.001)$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)



(a) Phase portrait (b) Equilibrium dist

(c) Mean trajectory (d) Normalized ACF

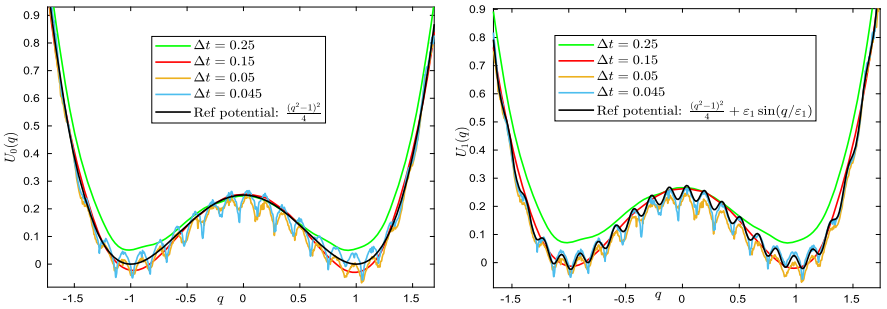
Fig. 3. Fig (a): Comparison on the surrogate Hamiltonian (4.1) via $U_0(q)$ with $\delta = 0.1$ and via $V(q)$ with $h = 5e-4$. **Fig (b)-(d):** Comparison on the surrogate Langevin (4.2) via $U_0(q)$ with $\delta = 0.1$ and via $V(q)$ with $h = 5e-4$. $V(q) = \frac{q^2}{2} + \varepsilon_1 \sin(q/\varepsilon_1) + \varepsilon_2 \sin(q/\varepsilon_2)$. Parameters are $(\varepsilon_1, \varepsilon_2) = (0.05, 0.001)$ and $T = 50$, initial distributions for $(q, p) \sim \mathcal{N}(-2, 1)$.

$$V(q) = (q-1)^2 (q+1)^2/4 + \varepsilon_1 \sin(q/\varepsilon_1) + \varepsilon_2 \sin(q/\varepsilon_2) \text{ with } (\varepsilon_1, \varepsilon_2) = (0.025, 0.001). \tag{4.6}$$

Similar to the quadratic example, we learnt $U_k(q)$ at each scale by choosing different time-stepping size δ. The results of learnt U_k with respect to various step sizes are plotted in Fig. 4. We do capture scales of ε_0 and ε_1 by employing suitable δ for (3.2). We next compare the performances of surrogate models via U_0 verse $V(q)$ for both Hamiltonian and Langevin. The results are summarized in Fig. 5. We still observe very good match from the plots.

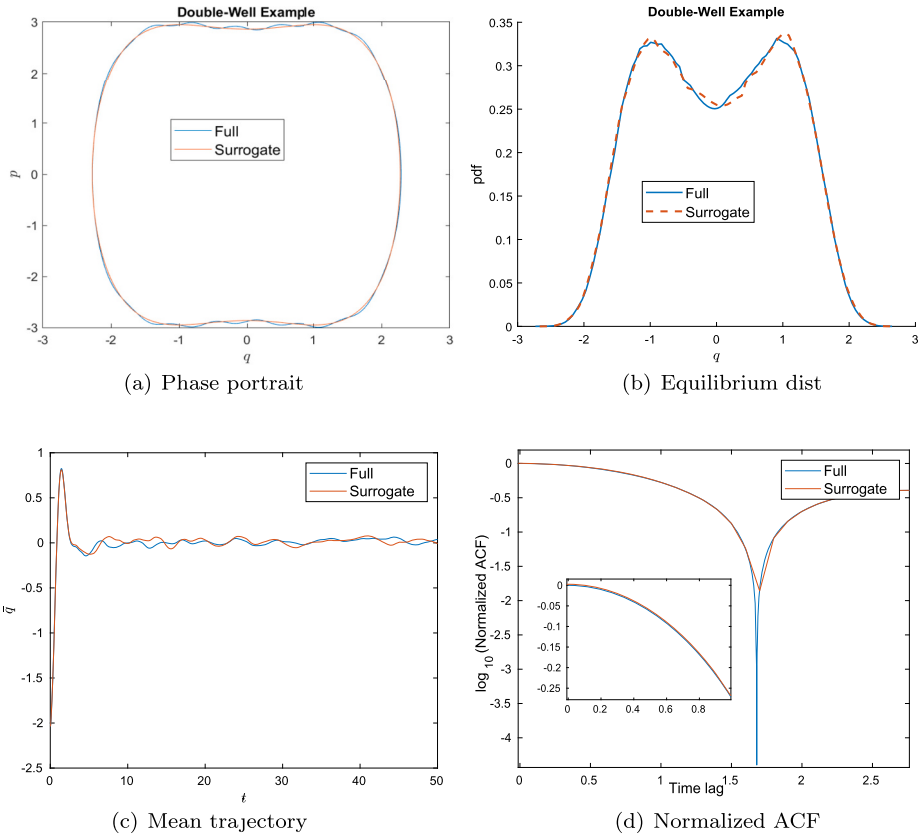
4.3. 1D example with lots of scales but no clear separation between adjacent scales

Practical problems are oftentimes multiscale but without a sharp scale separation (two famous examples are fluid and molecular dynamics). Typically there is a wide range of scales, where adjacent two scales are not well separated but the largest and smallest scale are very well separated. To consider a simplified version where the focus is just to decompose a multiscale function, we



(a) Reference $\frac{(q^2-1)^2}{4}$ (b) Reference $\frac{(q^2-1)^2}{4} + \varepsilon_1 \sin(q/\varepsilon_1)$

Fig. 4. The plots of learnt functions from simulated data with various time step sizes δ and the reference function $U_k(q)$. The exact function of full scales is $V(q) = \frac{(q^2-1)^2}{4} + \varepsilon_1 \sin(q/\varepsilon_1) + \varepsilon_2 \sin(q/\varepsilon_2)$. Parameters are $(\varepsilon_1, \varepsilon_2) = (0.025, 0.001)$.



(a) Phase portrait (b) Equilibrium dist

(c) Mean trajectory (d) Normalized ACF

Fig. 5. Fig (a): Comparison on the surrogate Hamiltonian (4.1) via $U_0(q)$ with $\delta = 0.1$ and via $V(q)$ with $h = 5e-4$. **Fig (b)-(d):** Comparison on the surrogate Langevin (4.2) via $U_0(q)$ with $\delta = 0.1$ and via $V(q)$ with $h = 5e-4$. $V(q) = (q^2-1)^2/4 + \varepsilon_1 \sin(q/\varepsilon_1) + \varepsilon_2 \sin(q/\varepsilon_2)$ with $(\varepsilon_1, \varepsilon_2) = (0.025, 0.001)$. $T = 50$, initial distributions for $q \sim \mathcal{N}(-2, 1)$ and $p \sim \mathcal{N}(0, 1)$.

investigate again problem (1.1), however with $1 \gtrsim \varepsilon_1 \gtrsim \cdots \gtrsim \varepsilon_K$ and $1 \gg \varepsilon_K$. Suppose δ is chosen at scale k, i.e. $\delta \approx \varepsilon_k$, the effective function should not be exactly $V_0(q) + \sum_{j=1}^{k} V_{j,\varepsilon_j}$, but with some additional correction as percolated from nearby smaller scales ε_{k+1}, etc. However, there is no clear or unique theoretical justification on how this correction works. We would like to inspect what our approach would produce in this case, and how the corresponding surrogate model approximates the original full model.

For this purpose, we consider a toy test problem

$$V(q) = \frac{(q - \pi/2)^2}{4} + \sum_{i=1}^{N} \frac{\cos(i^2 \times q)}{i^2}. \tag{4.7}$$

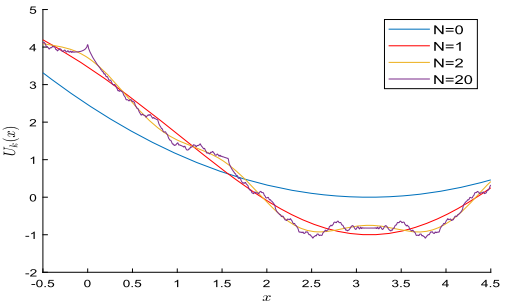
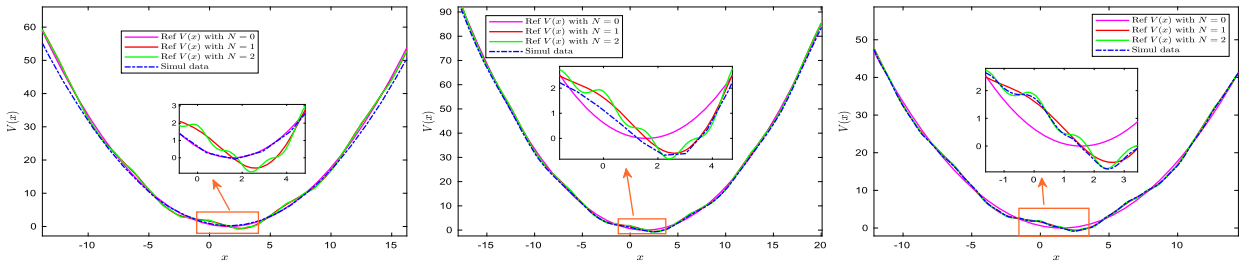


Fig. 6. Zoomed-in plot of $V(q) = \frac{(q-\pi/2)^2}{4} + \sum_{i=1}^{N} \frac{\cos(i^2 q)}{i^2}$ with different values of N.



(a) $\delta = 0.6500$ (b) $\delta = 0.2601$ (c) $\delta = 0.1309$

Fig. 7. Learning potentials $U_k(q)$ at different scales. The exact potential is $V(q) := \frac{(q-\pi/2)^2}{4} + \sum_{i=1}^{20} \cos(i^2 \times q)/i^2$.

A demonstration of the exact potential at different scales can be found in Fig. 6. Notice that different i's (besides the 1st term) only nominally correspond to different scales. While adjacent scales are close to each other, when N is large, $V(\cdot)$ clearly exhibits a wide range of well separated scales. In addition, one cannot directly read off $\{\varepsilon_k\}$ and $U_k(q)$ from the expression of $V(q)$, because different i's can contribute to the same scale (and other scales too) due to a lack of clear scale separation. Instead, it is reasonable to postulate the existence of some hidden functions that represent different scales.

To learn the mesocale $\{U_k(q)\}$ at various scales, we fix $\gamma = 0.5$, mass $M = 1$ and choose the total number of time steps to be $N_t = 5 \times 10^9$. We find several scales of potential functions via tuning the step sizes δ. The results are present in Fig. 7. We note from simulation that there is a macroscopic scale ε_0 around $\varepsilon_0 \sim \delta = 0.65$ as the learnt function is close to the referenced $N = 0$. The two following scales corresponding to $N = 1$ and $N = 2$ are around $\varepsilon_1 \sim \delta = 0.2601$ and $\varepsilon_2 \sim \delta = 0.1309$. The numerical separation between ε_1 and ε_2 is not as clear as those at ε_0 (i.e., $N = 0$). In fact, ε_1 and ε_2 are just weakly separated due to the narrow value gap in $V(q)$ between $N = 1$ and $N = 2$. Moreover, as demonstrated in Fig. 8, the accuracy of surrogate Hamiltonian and Langevin models constructed by our learning algorithm is better than those constructed by a simple truncation of objective function due to the mixed scales among the truncation.

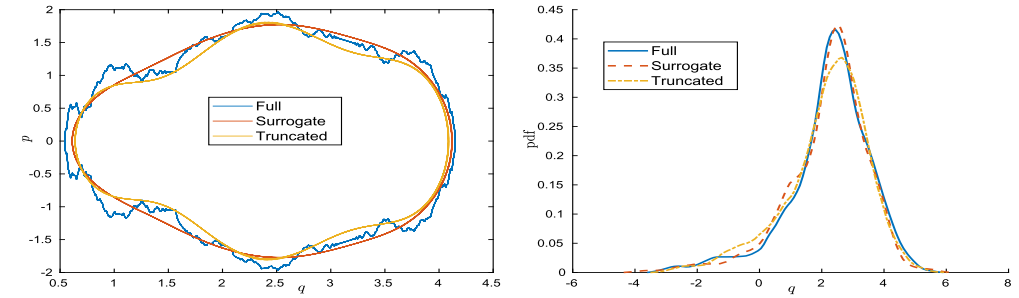
4.4. 2D quadratic function with anisotropic small scale

We now consider multi-dimensional examples. Throughout these examples, we will assume zero knowledge about how V decomposes in scales even though we wrote down its expression analytically.

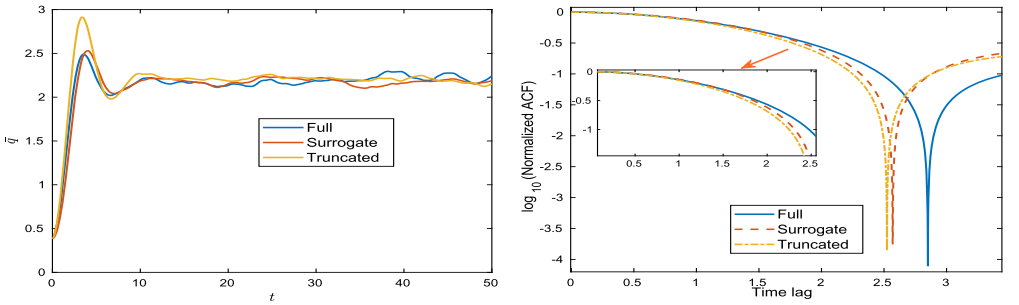
The first one is a two-dimensional quadratic potential, $\boldsymbol{q} = (x, y)$ with one known anisotropic small scale

$$\begin{aligned} V(x,y) &= V_0 + V_1 \\ &= \frac{1}{4}(2x + y - 1)^2 + (x - y - 1)^2 + \varepsilon\left(\sin(x/\varepsilon) + \sin((x+y)/\varepsilon)\right), \quad \varepsilon = 10^{-5}. \end{aligned} \tag{4.8}$$

We apply the two-stage Algorithm 1 with $\delta = 0.01$ and $N_t = 2 \times 10^7$ to learn the macroscopic potential $U_0 = V_0(x, y)$. During stage 1, the scalar friction γ is chosen to be $\gamma = 0.15$. During stage 2, we estimate the variance of small scale contributions and set Γ accordingly. We then compare the learning function from using γI only with that from the two-stage algorithm. The results are summarized in Fig. 9. We emphasize the mismatch of green trajectory data is generated without normalizing the variance of small scale, whereas the data from two-stage simulation can capture the macroscopic U_0 very well.

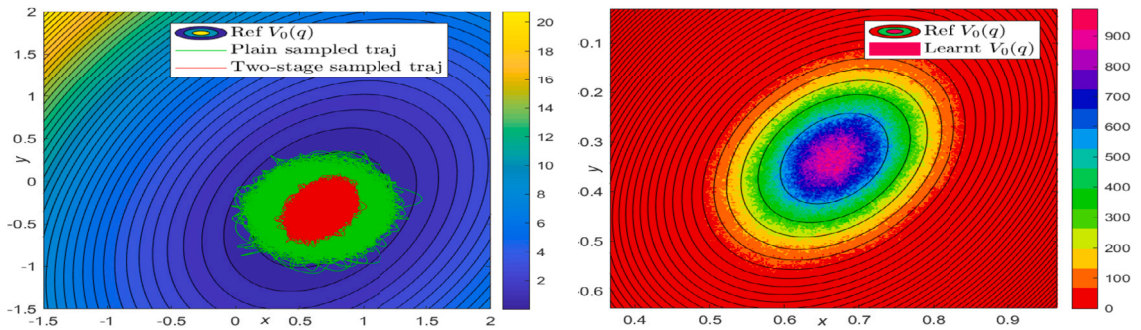


(a) Phase portrait (b) Equilibrium dist $\pi_\delta(q)$



(c) Mean trajectory (d) Normalized ACF

Fig. 8. Fig (a): Comparison on the surrogate Hamiltonian (4.1) via $U_0(q)$ with $\delta = 0.1309$ and via $V(q)$ with $h = 1e-3$. **Fig (b)-(d):** Comparison on the surrogate Langevin (4.2) via $U_0(q)$ with $\delta = 0.1309$ and via $V(q)$ (4.7) for $N = 20$ as well as the truncated version $N = 2$ with $h = 1e-3$. The initial distributions for $q \sim \mathcal{U}(-1/2, 1/2) + 0.38$ and $p \sim \mathcal{U}(-1/2, 1/2)$ with $\mathcal{U}$ presenting uniform distribution.



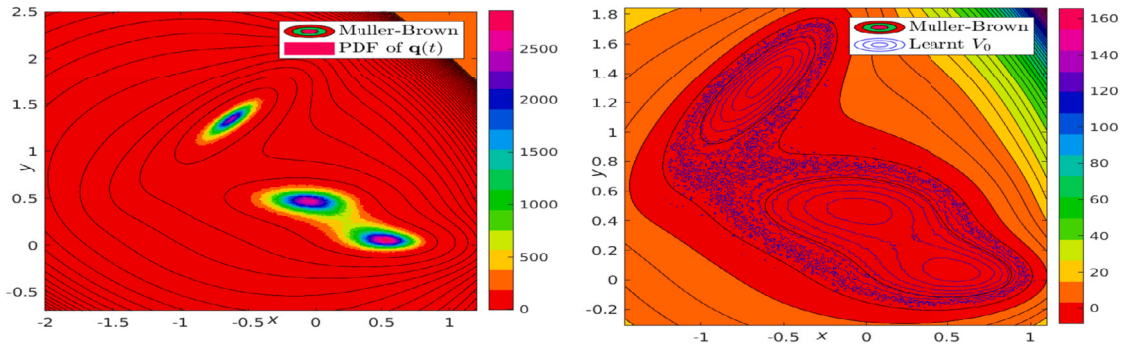
(a) Ref V_0 and traj of $\boldsymbol{q}(t)$ (b) Learning results U_0

Fig. 9. Learning results of macroscopic U_0. The exact potential is given in (4.8) and the step size is $\delta = 0.01$. **Left:** the green data represents trajectory generated via the scalar friction γI with $\gamma = 0.15$ only; and the red data represents trajectory generated via the two-stage Algorithm 1. **Right:** zoomed-in view of the two-stage learning potential results.

4.5. 2D Müller-Brown potential with one known anisotropic small scale

Müller-Brown potential is a common toy model used in molecular dynamics research, due to its high nonlinearity and the existence of multiple minima. We now consider the macrosopic potential to be a modification of Müller-Brown potential, and the microscopic scale to be some anisotropic toy function. More precisely, consider

$$\begin{aligned} V(x,y) &= V_0(x,y) + V_1(x,y), \\ V_0(x,y) &= 0.1 \times \Big(V_q(x,y) + V_m(x,y) \Big), \\ V_1(x,y) &= \varepsilon \Big(\sin\big(x/\varepsilon\big) + \sin\big((-x+y)/\varepsilon\big) \Big), \quad \varepsilon = 10^{-5}, \end{aligned} \tag{4.9}$$



(a) Ref V_0 and PDF of $\boldsymbol{q}$ (b) Learning results U_0

Fig. 10. Learning results of macroscopic U_0 at three wells. The exact potential is given in (4.9) and the simulation step size is $\delta = 0.05163$. The three wells of $U_0(x, y) = V_0$ are learnt simultaneously.

where $V_q(x, y)$ is a simple quadratic potential $V_q(x, y) = 35.0136(x - x_c^*)^2 + 59.8399(y - y_c^*)^2$, and $V_m(x, y)$ denotes the Müller-Brown potential

$$\begin{aligned} V_m(x,y) = & -200\exp\big(-(x-1)^2 - 10y^2\big) - 100\exp\big(-x^2 - 10(y-0.5)^2\big) \\ & - 170\exp\big(-6.5(x+0.5)^2 + 11(x+0.5)(y-1.5) - 6.5(y-1.5)^2\big) \\ & + 15\exp\big(0.7(x+1)^2 + 0.6(x+1)(y-1) + 0.7(y-1)^2\big). \end{aligned}$$

Here, (x_c^*, y_c^*) denotes the center of the middle well of V_m. Notice that $V_q(x, y)$ is introduced so that the depths of all three wells of the modified Müller-Brown (MB) function are better leveled to each order. This is because, the original Müller-Brown without modification has three local minima with very different function values, and therefore at the thermal equilibrium (i.e. Gibbs distribution) some potential well is exponentially less likely visited than the others, which both reduces the learning accuracy of the corresponding mode and makes visualization difficult.

We apply the two-stage Algorithm with $\delta = 0.05163$ and $N_t = 1 \times 10^8$ to learn the macroscopic scale $U_0 = V_0(x, y)$. In the stage 1, the scalar friction γ is chosen to be $\gamma = 0.014$. The results are summarized in Fig. 10. Note that we can learn all three energy wells at the same time with the existence of anisotropic small scale components.

5. Conclusion, discussion, and future possibilities

In this study, we introduced a novel algorithm for decomposing and learning every scale of a given multiscale objective function in multi-dimensions. Our method leverages the controllable algorithmic implicit bias inherent in the stochasticity of large learning rate gradient descent, thus facilitating the automatic generation of distinct data sets across varying scales. Upon successfully decomposing the multiscale objective function, many applications can be enabled. As a demonstration, we constructed a surrogate model aligned with the equilibrium distribution and dynamic mean path and auto-correlation of data produced by the original objective function at the scale of interest.

Additionally, we devised a two-stage algorithm to segregate variables of interest within an anisotropic stochastic forcing environment. By adjusting the system's mass tensor, we estimated the covariance of the stochastic forcing, and an appropriate friction matrix could be chosen to bring the equilibrium distribution back towards a Gibbs distribution, which possesses an analytical expression.

Meanwhile, the proposed strategy has certain limitations. The first limitation is the statistical accuracy due to finite samples. Detailed choices for the function fitting also matter; for example, the effective potential can only be resolved to the scale dictated by the bin width, and which (parameterized) model to use for the regression makes a difference. Although these are all generic and standard problems, we would still like to point them out, because a consequence is, the current implementation requires a large amount of (self-generated) data.

The second limitation is that the performance of the proposed method depends on various hyperparameters. For instance, to learn the objective function U_k at scale ε_k, the step size (learning rate) δ needs to be around ε_k. When the scale separation is large, i.e., $\varepsilon_k \gg \varepsilon_{k+1}$, the choice of δ is relatively flexible as demonstrated in examples of Subsection 4.2, 4.4 and 4.5. When the scale separation is narrow, the tuning of δ becomes more subtle as demonstrated in Subsection 4.3 and Fig. 7. The data size also needs to be sufficiently large to reduce statistical errors from simulation and estimation. The artificial mass tensor M needs to be chosen to balance the numerical stability and the efficiency of sampling. The empirically tuned M, as suggested in Remark 3.5, appears to work well in our experiments.

Moreover, the selection of hyperparameters and the global learning performance of the proposed method depend on the energy barriers of the macroscopic objective function. When there are no energy barriers or the energy barriers are of order $O(1)$ in V_0, we typically select a mass tensor M with entries $M_{ij} \in [1/2, 1]$, set $\hat{\beta} = 1$, and choose a suitable step size δ to achieve good global

learning results for a wide range of initial conditions $(\boldsymbol{q}_0, \boldsymbol{p}_0)$. However, when the energy barriers are large, it may be nontrivial to find hyperparameter values and a simulation step size δ for effectively learning the objective function globally.

Possible future directions include 1) to explore more statistical tools to have better estimation of π_δ and more robust regression of $U_k(\boldsymbol{q})$, and 2) to derive *a priori* error estimate of learning U_k in terms of various problem and hyper parameters settings.

CRediT authorship contribution statement

Xingjie Helen Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Molei Tao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

X. Li's is grateful for partial support by the NSF Award DMS-1847770 and the 2023-2024 UNC Charlotte faculty research grant. M. Tao is grateful for partial support by the NSF Award DMS-1847802, by Cullen-Peck Scholar Award (Georgia Institute of Technology), and Emory-GT AI. Humanity Award (Emory University and Georgia Institute of Technology).

References

[1] Assyr Abdulle, Andrea Di Blasio, A Bayesian numerical homogenization method for elliptic multiscale inverse problems, SIAM/ASA J. Uncertain. Quantificat. 8 (1) (2020) 414–450.
[2] Assyr Abdulle, Weinan E, Bjorn Engquist, Eric Vanden-Eijnden, The heterogeneous multiscale method, Acta Numer. 21 (2012) 1–87.
[3] Assyr Abdulle, Giacomo Garegnani, Grigorios A. Pavliotis, Andrew M. Stuart, Andrea Zanoni, Drift estimation of multiscale diffusions based on filtered data, Found. Comput. Math. (2021) 1–52.
[4] Assyr Abdulle, Grigorios A. Pavliotis, Andrea Zanoni, Eigenfunction martingale estimating functions and filtered data for drift estimation of discretely observed multiscale diffusions, Stat. Comput. 32 (2) (2022) 34.
[5] Michael Aizenman, Simone Warzel, Random Operators, vol. 168, American Mathematical Soc., 2015.
[6] Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, et al., Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences, npj Digit. Med. 2 (2019) 1–11.
[7] Grégoire Allaire, Robert Brizzi, A multiscale finite element method for numerical homogenization, Multiscale Model. Simul. 4 (3) (2005) 790–812.
[8] Sabine Attinger, Petros D. Koumoutsakos, Multiscale Modelling and Simulation, Springer, 2004.
[9] Ivo Babuška, Manil Suri, The p- and hp versions of the finite element method, an overview, Comput. Methods Appl. Mech. Eng. 80 (1–3) (1990) 5–26.
[10] Alain Bensoussan, Jacques-Louis Lions, George Papanicolaou, Asymptotic Analysis for Periodic Structures, vol. 374, American Mathematical Soc., 2011.
[11] Xavier Blanc, Claude Le Bris, Pierre-Louis Lions, Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques, C. R. Math. 343 (11–12) (2006) 717–724.
[12] James H. Bramble, Joseph E. Pasciak, Jinchao Xu, The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms, Math. Comput. 56 (193) (1991) 1–34.
[13] James H. Bramble, Xuejun Zhang, The analysis of multigrid methods, Handb. Numer. Anal. 7 (2000) 173–415.
[14] Alexandre Chorin, Panagiotis Stinis, Problem reduction, renormalization, and memory, Commun. Appl. Math. Comput. Sci. 1 (1) (2007) 1–27.
[15] K.C. Chou, A.S. Willsky, R. Nikoukhah, Multiscale systems, Kalman filters, and Riccati equations, IEEE Trans. Autom. Control 39 (3) (1994) 479–492.
[16] Bernardo Cockburn, Johnny Guzmán, See-Chew Soon, Henry K. Stolarski, An analysis of the embedded discontinuous Galerkin method for second-order elliptic problems, SIAM J. Numer. Anal. 47 (4) (2009) 2686–2707.
[17] L. Cohen, Time–Frequency Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1995.
[18] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, PA, 1992.
[19] Alain de Cheveigné, Israel Nelken, Filters: when, why, and how (not) to use them, Neuron 102 (2) (2019) 280–293.
[20] E. De Giorgi, Sulla convergenza di alcune successioni d'integrali del tipo dell'area, Ennio De Giorgi 414 (1975) 64.
[21] Daniel Dylewsky, Molei Tao, J. Nathan Kutz, Dynamic mode decomposition for multiscale nonlinear physics, Phys. Rev. E 99 (6) (2019) 063311.
[22] Weinan E, Bjorn Engquist, The heterogeneous multi-scale methods, Commun. Math. Sci. 1 (2003) 87–133.
[23] Yalchin Efendiev, Thomas Y. Hou, Multiscale Finite Element Methods: Theory and Applications, vol. 4, Springer Science & Business Media, 2009.
[24] Bjorn Engquist, Yen-Hsi Tsai, Heterogeneous multiscale methods for stiff ordinary differential equations, Math. Comput. 74 (252) (2005) 1707–1742.
[25] Zhipeng Feng, Dong Zhang, Ming J. Zuo, Adaptive mode decomposition methods and their applications in signal analysis for machinery fault diagnosis: a review with examples, IEEE Access 5 (2017) 24301–24331.
[26] Mark G. Frei, Osorio Ivan, Intrinsic time-scale decomposition: time–frequency–energy analysis and real-time filtering of non-stationary signals, Proc. R. Soc. A 463 (2007) 321–342.
[27] Dietmar Gallistl, Timo Sprekeler, Endre Süli, Mixed finite element approximation of periodic Hamilton–Jacobi–Bellman problems with application to numerical homogenization, Multiscale Model. Simul. 19 (2) (2021) 1041–1065.

[28] Alexander N. Gorban, Nikolaos K. Kazantzis, Ioannis G. Kevrekidis, Hans Christian Öttinger, Constantinos Theodoropoulos, Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena, Springer, 2006.
[29] Mohinder S. Grewal, Angus P. Andrews, Kalman Filtering: Theory and Practice, Prentice Hall, 1993.
[30] Michael Griebel, Christian Rieger, Barbara Zwicknagl, Multiscale approximation and reproducing kernel Hilbert space methods, SIAM J. Numer. Anal. 53 (2) (2015) 852–873.
[31] A. Haar, Zur theorie der orthogonalen funktionensysteme, Math. Ann. 69 (1910) 331–371.
[32] Jiequn Han, Chao Ma, Zheng Ma, Weinan E, Uniformly accurate machine learning-based hydrodynamic models for kinetic equations, Proc. Natl. Acad. Sci. 116 (44) (2019) 21983–21991.
[33] Jan S. Hesthaven, Gianluigi Rozza, Benjamin Stamm, et al., Certified Reduced Basis Methods for Parametrized Partial Differential Equations, vol. 590, Springer, 2016.
[34] Thomas Y. Hou, Xiao-Hui Wu, A multiscale finite element method for elliptic problems in composite materials and porous media, J. Comput. Phys. 134 (1) (1997) 169–189.
[35] Paul Houston, Bill Senior, Endre Süli, Sobolev regularity estimation for hp-adaptive finite element methods, in: Numerical Mathematics and Advanced Applications: Proceedings of ENUMATH 2001 the 4th European Conference on Numerical Mathematics and Advanced Applications Ischia, July 2001, Springer, 2003, pp. 631–656.
[36] Thomas J.R. Hughes, Guglielmo Scovazzi, Pavel B. Bochev, Annalisa Buffa, A multiscale discontinuous Galerkin method with the computational structure of a continuous Galerkin method, Comput. Methods Appl. Mech. Eng. 195 (19–22) (2006) 2761–2787.
[37] Shi Jin, Asymptotic-preserving schemes for multiscale physical problems, Acta Numer. 31 (2022) 415–489.
[38] Shi Jin, Zheng Ma, Keke Wu, Asymptotic-preserving neural networks for multiscale time-dependent linear transport equations, J. Sci. Comput. 94 (3) (2023) 57.
[39] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, Liu Yang, Physics-informed machine learning, Nat. Rev. Phys. 3 (6) (2021) 422–440.
[40] Mathieu Kessler, Michael Sørensen, Estimating equations based on eigenfunctions for a discretely observed diffusion process, Bernoulli 5 (2) (1999) 299–314.
[41] Ioannis G. Kevrekidis, C. William Gear, Gerhard Hummer, Equation-free: the computer-aided analysis of complex multiscale systems, AIChE J. 50 (7) (2004) 1346–1355.
[42] Ioannis G. Kevrekidis, C. William Gear, James M. Hyman, Panagiotis G. Kevrekidis, Olof Runborg, Constantinos Theodoropoulos, Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis, Commun. Math. Sci. 1 (4) (2003) 715–762.
[43] Ioannis G. Kevrekidis, Giovanni Samaey, Equation-free multiscale computation: algorithms and applications, Annu. Rev. Phys. Chem. 60 (2009) 321–344.
[44] Lingkai Kong, Molei Tao, Stochasticity of deterministic gradient descent: large learning rate for multiscale objective function, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 2625–2638.
[45] Sergei Mikhailovich Kozlov, Averaging of random operators, Mat. Sb. 151 (2) (1979) 188–202.
[46] S. Krumscheid, G.A. Pavliotis, S. Kalliadasis, Semiparametric drift and diffusion estimation for multiscale diffusions, Multiscale Model. Simul. 11 (2) (2013) 442–473.
[47] J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, Joshua L. Proctor, Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, SIAM, 2016.
[48] Jacques Laskar, Frequency analysis for multi-dimensional systems. Global dynamics and diffusion, Phys. D: Nonlinear Phenom. 67 (1–3) (1993) 257–281.
[49] T. Lindeberg, Scale-Space Theory in Computer Vision, Kluwer Academic Publishers, Netherlands, 1994.
[50] J-L. Lions, Résolution d'edp par un schéma en temps «pararéel» (A "parareal" in time discretization of pde's), Acad. Sci. Paris C. R. Ser. Sci. Math. 332 (7) (2001) 661–668.
[51] Andrew J. Majda, John Harlim, Filtering Complex Turbulent Systems, Cambridge University Press, 2012.
[52] François Murat, Luc Tartar, H-Convergence: Topics in the Mathematical Modelling of Composite Materials, 2018, pp. 21–43.
[53] Roland Opfer, Tight frame expansions of multiscale reproducing kernels in Sobolev spaces, Appl. Comput. Harmon. Anal. 20 (3) (2006) 357–374.
[54] A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[55] Owhadi Houman, Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games, SIAM Rev. 59 (1) (2017) 99–149.
[56] Houman Owhadi, Clint Scovel, Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design, vol. 35, Cambridge University Press, 2019.
[57] George C. Papanicolaou, Boundary value problems with rapidly oscillating random coefficients, Colloq. Math. Soc. János Bolyai 27 (1979) 853–873.
[58] G.A. Pavliotis, A.M. Stuart, Parameter estimation for multiscale diffusions, J. Stat. Phys. 127 (4) (2007).
[59] Grigorios A. Pavliotis, Stochastic Processes and Applications, Springer, 2014.
[60] Grigorios A. Pavliotis, Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations, Texts in Applied Mathematics, vol. 60, Springer, New York, 2014.
[61] Grigoris Pavliotis, Andrew Stuart, Multiscale Methods: Averaging and Homogenization, Springer Science & Business Media, 2008.
[62] Grace C.Y. Peng, Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Savador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, et al., Multiscale modeling meets machine learning: what can we learn?, Arch. Comput. Methods Eng. 28 (3) (2021) 1017–1037.
[63] Alfio Quarteroni, Gianluigi Rozza, et al., Reduced Order Methods for Modeling and Computational Reduction, vol. 9, Springer, 2014.
[64] Peter J. Schmid, Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech. 656 (2010) 5–28.
[65] Sergio Spagnolo, Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche, Ann. Sc. Norm. Super. Pisa, Cl. Sci. 22 (4) (1968) 571–597.
[66] G. Strang, T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, Wellesley, MA, 1996.
[67] Gilbert Strang, Wavelets and dilation equations: a brief introduction, SIAM Rev. 31 (4) (1989) 614–627.
[68] Theofanis Strouboulis, Ivo Babuška, Kevin Copps, The design and analysis of the generalized finite element method, Comput. Methods Appl. Mech. Eng. 181 (1–3) (2000) 43–69.
[69] Molei Tao, Houman Owhadi, Jerrold E. Marsden, Nonintrusive and structure preserving multiscale integration of stiff odes, sdes, and Hamiltonian systems with hidden slow dynamics via flow averaging, Multiscale Model. Simul. 8 (4) (2010) 1269–1324.
[70] Molei Tao, Houman Owhadi, Jerrold E. Marsden, Space-time FLAVORS: finite difference, multisymlectic, and pseudospectral integrators for multiscale PDEs, Dyn. Partial Differ. Equ. 8 (1) (2011) 21–46.
[71] Henry C. Thode, Testing for Normality, vol. 164, CRC Press, 2002.
[72] Wei Wang, Xiantao Li, Chi-Wang Shu, The discontinuous Galerkin method for the multiscale modeling of dynamics of crystalline solids, Multiscale Model. Simul. 7 (1) (2008) 294–320.
[73] Weinan E, Principles of Multiscale Modeling, Cambridge University Press, 2011.
[74] Weinan E, Bjorn Engquist, Xiantao Li, Weiqing Ren, Eric Vanden-Eijnden, Heterogeneous multiscale methods: a review, Commun. Comput. Phys. 2 (3) (2007) 367–450.
[75] Linfeng Zhang, Han Wang, Weinan E, Reinforced dynamics for enhanced sampling in large atomic and molecular systems, J. Chem. Phys. 148 (2018) 124113.
[76] Zhiqiang Zhou, Bo Wang, Sun Li, Mingjie Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters, Inf. Fusion 30 (2016) 15–26.