



Self-Test Loss Functions for Learning Weak-Form Operators and Gradient Flows

Yuan Gao¹ · Quanjun Lang² · Fei Lu³ 

Received: 3 January 2025 / Revised: 13 December 2025 / Accepted: 23 May 2026 /

Published online: 5 June 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

The construction of loss functions presents a major challenge in data-driven modeling involving weak-form operators in PDEs and gradient flows, particularly due to the need to select test functions appropriately. We address this challenge by introducing self-test loss functions, which employ test functions that depend on the unknown parameters, specifically for cases where the operator depends linearly on the unknowns. The proposed self-test loss function conserves energy for gradient flows and coincides with the expected log-likelihood ratio for stochastic differential equations. Importantly, it is quadratic, facilitating theoretical analysis of identifiability and well-posedness of the inverse problem, while also leading to efficient parametric or nonparametric regression algorithms. It is computationally simple, requiring only low-order derivatives or even being entirely derivative-free, and numerical experiments demonstrate its robustness against noisy and discrete data.

Keywords Inverse problem · Weak form · Variational inference · Gradient flow · Operator learning

Mathematics Subject Classification 35R30 · 68T09

Yuan Gao was partially supported by NSF under Award DMS-2204288. Fei Lu was partially supported by NSF DMS-2238486.

✉ Fei Lu
feilu@math.jhu.edu

Yuan Gao
gao662@purdue.edu

Quanjun Lang
quanjun.lang@duke.edu

¹ Department of Mathematics, Purdue University, West Lafayette, IN, USA

² Department of Mathematics, Duke University, Durham, NC, USA

³ Department of Mathematics, Johns Hopkins University, Baltimore, MD, USA

1 Introduction

Learning governing equations from data is a fundamental task in many areas of science and engineering, such as physics, biology, and geosciences [6, 14, 15, 25, 35, 37, 41]. The governing equations allow us to model complex systems, predict future behavior, and develop effective control strategies. They are often in the form of partial differential equations (PDEs), such as gradient flows [3, 11, 12, 19] and diffusion models [4, 13, 18, 30, 42, 46]. To learn these equations, it is necessary to use data to approximate the differential operators. However, real-world data is often *noisy and discrete*, leading to large errors in derivative approximations and unreliable estimators when using strong-form equations.

Weak-form equations provide a more versatile framework. By using smooth test functions with integration by parts, weak forms use lower-order differential operators, thereby offering improved robustness to noisy and discrete data [10, 13, 21, 31, 32, 43, 48].

However, constructing loss functions for variational inference of weak-form equations poses a major challenge. This difficulty arises because the weak form requires test functions to be dense in the dual space, typically an infinite-dimensional function space. In classical approaches, test functions are often chosen to be smooth and compactly supported, with Galerkin basis functions being a prominent example [32]. These methods are often limited to low-dimensional problems and are not scalable to high-dimensional settings, such as the Wasserstein gradient flows of probability measures in high-dimensional spaces. Importantly, since universal test functions are agnostic to the data and the model, it is necessary to use a large set of such test functions to ensure that all relevant information from the data is captured, which often leads to redundancy and computational inefficiency.

We address this challenge by introducing *self-test loss functions* for cases where the operator depends linearly on the (function-valued) parameter. The key idea is to employ test functions that depend on the unknown parameter itself and the data, which we call *self-testing functions*. Such test functions are automatically determined by the operator and the data. Thus, they automate the construction of the loss function.

The proposed loss function is suitable for various weak-form operators, including the high-dimensional gradient flows and diffusion models. In particular, the self-test loss function is quadratic. It facilitates theoretical analysis of identifiability and well-posedness of the inverse problem. It also enables efficient parametric and nonparametric regression algorithms. It is computationally simple, requiring only low-order derivatives or even being entirely derivative-free. Our numerical experiments demonstrate its robustness against noisy and discrete data.

1.1 Problem Settings and Main Results

Consider the problem of estimating the (function-valued) parameter ϕ in the operator $R_\phi : \mathbb{X} \rightarrow \mathbb{Y}$ in the weak-form equation:

$$R_\phi[u] = f \Leftrightarrow \langle R_\phi[u], v \rangle = \langle f, v \rangle, \quad \forall v \in \mathbb{Y}^* \quad (1.1)$$

from data consisting of noisy discrete observations of input-output pairs:

$$\mathcal{D} = \{(u_l, f_l)\}_{l=1}^L. \quad (1.2)$$

Here, $\mathbb{X}, \mathbb{Y}$ are metric spaces, $\mathbb{Y}^*$ is the dual space of $\mathbb{Y}$, and $\langle \cdot, \cdot \rangle$ means the dual pair between $\mathbb{Y}$ and $\mathbb{Y}^*$. The operator $R_\phi : \mathbb{X} \rightarrow \mathbb{Y}$ can be either linear or nonlinear. Depending

on the operators, the data can be the functions at discrete spatial-time meshes or empirical distributions of samples; see (1.6), (1.9) and (1.11) below.

We assume that the operator $R_\phi[u]$ depends linearly on ϕ when u is fixed, that is,

$$R_{\alpha\phi_1+\beta\phi_2}[u] = \alpha R_{\phi_1}[u] + \beta R_{\phi_2}[u], \quad (1.3)$$

for any $\alpha, \beta \in \mathbb{R}$ and any function ϕ_1 and ϕ_2 such that the operator is well-defined. We assume no prior knowledge of ϕ , except that the operator $R_\phi[u]$ is well-defined.

To construct a loss function using the weak form equation, we introduce *self-testing functions* $v_\phi[u] \in \mathbb{Y}^*$ defined from R_ϕ and u so that, for all ϕ, ψ ,

$$\langle R_\phi[u], v_\psi[u] \rangle = \langle R_\psi[u], v_\phi[u] \rangle \text{ (symmetry)}, \quad \langle R_\phi[u], v_\phi[u] \rangle \geq 0 \text{ (positivity)}. \quad (1.4)$$

The *self-test loss function* is

$$\mathcal{E}_D(\phi) = \frac{1}{L} \sum_{l=1}^L \langle R_\phi[u_l], v_\phi[u_l] \rangle - 2 \langle f_l, v_\phi[u_l] \rangle + C_0,$$

where C_0 is an arbitrary constant.

We demonstrate the self-test loss function in three settings involving function-valued parameters: Wasserstein gradient flows, a weak-form elliptic operator, and interacting particle systems with sequential ensembles of unlabeled data. Among these, Example 1 serves as a running example throughout the paper.

Example 1 (Wasserstein gradient flow) Estimate $\phi = (h, \Phi, V)$ in the Wasserstein gradient flow

$$\partial_t u = \nabla \cdot (u \nabla [v h'(u) + \Phi * u + V]) =: R_\phi[u], \quad (1.5)$$

where $h : \mathbb{R} \rightarrow \mathbb{R}$ is the diffusion rate function, $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is the pairwise interaction potential satisfying $\Phi(-x) = \Phi(x)$, and $V : \mathbb{R}^d \rightarrow \mathbb{R}$ is an external potential acting on each particle. The data consists of discrete noisy observations of solutions on a mesh $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$:

$$\mathcal{D}_1 = \{u_l(x_i)\}_{i,l=1}^{N,L}, \quad u_l(x_i) = u(x_i, t_l) + \epsilon_{i,l}, \quad (1.6)$$

where $\{\epsilon_{i,l}\}$ are noises or measurement errors. The self-testing function is $v_\phi[u] = v h'(u) + \Phi * u + V$, and the self-test loss function is

$$\mathcal{E}_u(\phi) := \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} [u |\nabla [v h'(u) + \Phi * u + V]|^2 - 2 \partial_t u [v h'(u) + \Phi * u + V]] dx dt, \quad (1.7)$$

see Section 2.2 for a derivation and Section 4 for an analysis on identifiability and well-posedness of the inverse problem. One can construct an empirical loss function by approximating the integrals in (1.7) using data $\mathcal{D}_1$.

Example 2 (Weak-form operator) Estimate the coefficient $a : \mathbb{R}^d \rightarrow \mathbb{R}$ in the PDE:

$$R_a[u] := -\Delta(au) = f \quad (1.8)$$

from data consisting of discrete noisy observations on the spatial mesh $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$:

$$\mathcal{D}_2 = \{(u_l(x_i), f_l(x_i))\}_{i,l=1}^{N,L}. \quad (1.9)$$

The self-testing function is $v_a[u] = au$; see Section 2.3. Note that this inverse problem is different from the inverse conductivity problem (see e.g., [18]), where the goal is to estimate a in $\nabla \cdot (a \nabla u) = 0$ in Ω when given only $u|_{\partial\Omega} = f$.

Example 3 (Sequential ensembles of unlabeled data) Estimate the potentials $\Phi, V : \mathbb{R}^d \rightarrow \mathbb{R}$ in the differential equation of N -interacting particles,

$$\frac{d}{dt} X_t^i = -[\nabla V(X_t^i) + \frac{1}{N} \sum_{j=1}^N \nabla \Phi(X_t^i - X_t^j)], \quad 1 \leq i \leq N \quad (1.10)$$

from data consisting of M independent sequences of ensembles of **unlabeled** particles

$$\mathcal{D}_3 = \{(X_{t_l}^{i_l, (m)}, 1 \leq i_l \leq N)\}_{m, l=1}^{M, L}. \quad (1.11)$$

Since the particles are unlabeled, there is no information on their trajectories. Thus, the classical methods based on the derivatives $\frac{d}{dt} X_t^i$, see e.g., [27–29], are no longer applicable. We construct a loss function based on the weak-form equation of the empirical measures $\mu_N(x, t) := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}(x)$, and the self-testing function is $v_\phi[\mu_N] = \Phi * \mu_N + V$; see Section 2.4. Additionally, we demonstrate numerical estimation using neural network approximation in Section 5.3.

Key features of the self-test loss function. The quadratic self-test loss function conserves energy for gradient flows, aligns with the expected log-likelihood ratio for stochastic differential equations, facilitates theoretical analysis of identifiability and well-posedness, and leads to efficient parametric or nonparametric regression algorithms.

- It aims to match the energy dissipation for the Wasserstein gradient flow, and its minimizer *conserves the energy* of the data flow; see Theorem 3.2. The self-testing functions are the first variation of the free energy. Also, for the weak-form Fokker-Planck equation of the McKean-Vlasov stochastic differential equation (SDE), the self-test loss function coincides with the expectation of the negative log-likelihood ratio (see Theorem 3.3). As a result, a minimizer of the self-test loss function maximizes the expected likelihood.
- Importantly, the loss function is quadratic since both $R_\phi[u]$ and $v_\phi[u]$ are linear in ϕ . It facilitates the analysis on *identifiability* and *well-posedness* of the inverse problem based on the Hessian of the loss function. We demonstrate such an analysis for learning the diffusion rate function and the potentials in the Wasserstein gradient flow in Section 4.
- It also leads to computationally efficient parametric or nonparametric regression algorithms, using either least-squares or neural network regression. We demonstrate its robustness against noisy and discrete data in parametric and nonparametric estimations in Section 5.

1.2 Related Work

Weak formulations offer a robust and flexible foundation for addressing both forward and inverse PDE problems, and have thus attracted growing attention in recent years.

General forward and inverse problems using weak-form. For forward problems, machine learning methods rooted in variational principles include the Deep Ritz method [9], the Deep Galerkin method [40], variational physics-informed neural networks [7, 21, 22], and physics-informed graph neural Galerkin networks [10], among others. For inverse PDE problems, we refer to [18] and [13] for comprehensive overviews. In classical inverse settings, where data are often limited to boundary measurements (e.g., in the inverse conductivity problem) or spectral information (e.g., in inverse spectral problems), one must estimate both the solution and the unknown parameters simultaneously. Recent approaches, such as weak adversarial networks [2] and physics-informed graph neural Galerkin networks [10], use weak form equations to address these classical difficulties.

Table 1 Notations

Notations	Description
$R_\phi[\cdot]$, $v_\phi[\cdot]$	Operators $R_\phi[\cdot] : \mathbb{X} \rightarrow \mathbb{Y}$, $v_\phi[\cdot] : \mathbb{X} \rightarrow \mathbb{Y}^*$
ϕ	(Function-valued) parameter to be estimated from data
$\langle f, v \rangle$, $\langle \cdot, \cdot \rangle_H$	Dual operation with $f \in \mathbb{Y}$, $v \in \mathbb{Y}^*$; inner product in H
$\mathcal{E}(\cdot)$ $E_\phi(\cdot)$	$\mathbb{R}$ -valued loss function and energy function
$\mathcal{P}_2(\mathbb{R}^d)$	The space of probability measures with finite second moments

In contrast, our setting considers data consisting of PDE solutions sampled on discrete spatial grids or approximated by empirical measures, and the task is to estimate the PDE parameters. On the other hand, our self-test loss function can be applied to these methods, providing a systematic way to construct loss functions based on weak-form equations.

Regression based on weak-form. Weak-form methods for parameter estimation have been widely explored in sparse regression frameworks, including Weak-SINDy [31–33], Weak-PDE-LEARN [43], and other weak-form-based data-driven modeling approaches [4, 13, 36, 37, 41]. These methods rely on carefully designed families of smooth, compactly supported test functions that must be tailored to the data, domain, and PDE structure, a task that becomes increasingly challenging and computationally demanding in high dimensions. In contrast, the self-test loss function proposed in this work removes the need for such hand-crafted test sets by constructing test functions directly from the operator and the data in a canonical way. Moreover, it can be seamlessly integrated into the Weak SINDy framework: one may include the self-test function as an additional test function or augment the Weak SINDy loss with the self-test term, thereby combining SINDy’s sparsity-promoting structure with the robustness and adaptivity of the self-test formulation.

Energy variational approaches and gradient flow inference. Our framework is closely related to energy variational approaches [17, 30, 45] and gradient flow inference [4, 23]. The energy-dissipation-based loss [30, 45] shares conceptual similarities with the self-test loss function, aiming to preserve energy structures observed in the data. In particular, both approaches accommodate PDEs or stochastic differential equations for generalized diffusions and gradient flows, and can handle data defined on spatial grids or represented by particle samples. Furthermore, in the context of gradient flow inference, the self-test loss aligns with likelihood-based loss functions [23] and the quadratic loss [4]. By casting these methods into a unified variational inference framework, the self-test loss function extends their applicability beyond energy-dissipating systems to general weak formulations.

The rest of the paper is organized as follows. Section 2 defines the framework of the self-test loss functions and provides examples. In Section 3, we show that for general gradient flow, the self-test loss function’s minimizers conserve the energy. We also connect it with the likelihood of SDEs. In Section 4, we study the identifiability and well-posedness of the diffusion rate function and the potentials in aggregation-diffusion equations. We present numerical experiments in Section 5 and conclude in Section 6.

Notation Throughout the paper, we denote the true parameter by ϕ_* and observational data by f . We abuse the notation u , which may represent either a function $u(x)$ or $u(x, t)$ for a given t , as long as the context is clear. Table 1 lists the notations.

2 Self-Test Loss Functions

We first formulate the self-test loss function within a general *weak-form operator learning* setting that includes both weak-form PDEs and gradient flows. The formulation is then illustrated through the running examples. For clarity, we derive the loss functions under the assumption of continuous noiseless data, before discussing how they are approximated from discrete, noisy measurements in practice.

2.1 Weak-Form Operator Learning

The main idea behind the self-test loss function is to guide the minimization in the direction that explores the unknown parameter the most. Thus, we use the parameter to construct test functions.

Definition 1 (*Self-test loss function*) Consider the problems of estimating ϕ in the operator equation (1.1) from the dataset in (1.2), where the operator $R_\phi[u]$ is linear in ϕ . We call $v_\phi[u] \in \mathbb{Y}^*$ a **self-testing function** if it satisfies the self-testing properties:

$$\begin{aligned} \text{Symmetry:} \quad & \langle R_\phi[u], v_\psi[u] \rangle = \langle R_\psi[u], v_\phi[u] \rangle, \\ \text{Positivity:} \quad & \langle R_\phi[u], v_\phi[u] \rangle \geq 0, \\ \text{Linearity:} \quad & v_{\phi+\psi}[u] = v_\phi[u] + v_\psi[u], \end{aligned} \quad (2.1)$$

for any ϕ, ψ such that these operations are well-defined for all $u \in \{u_l\}_{l=1}^L$. We call

$$\mathcal{E}_D(\phi) = \sum_{l=1}^L \langle R_\phi[u_l], v_\phi[u_l] \rangle - 2 \langle f_l, v_\phi[u_l] \rangle + C_0 \quad (2.2)$$

a **self-test loss function**, where C_0 is an arbitrary constant.

The self-test loss function has three appealing properties. First, it is *quadratic* in the unknown parameter ϕ . Thus, it is convex, and its minimizers can be computed using the broad class of regression techniques. Also, the uniqueness of the minimizer can be established in a proper function space, as well as the well-posedness of the inverse problem; see Section 4. Second, it employs the weak form operator, which requires either a low-order derivative or no derivatives of u , thereby avoiding numerical errors when approximating derivatives from noisy discrete data. Lastly, in applications with probability gradient flow, it is particularly suitable for high-dimensional systems with ensemble data consisting of particle samples, as the loss function can be written as a combination of expectations; see Sections 2.4 and 3.2.

Two major tasks in the construction of the self-test loss function are (i) to find the self-testing function $v_\phi[u]$, and (ii) to select a proper parameter space for the minimization. Fortunately, the linearity of $R_\phi[u]$ in ϕ and the self-testing properties (2.1) provide clear clues on constructing $v_\phi[u]$. As examples, we explore such self-testing functions for weak-form PDEs and gradient flows in Sections 2.2–2.3. Meanwhile, the loss function indicates adaptive function spaces for the parameter, which we explore in Section 4.

The next proposition shows that any minimizer of the self-test loss function satisfies the weak-form equation when tested against all admissible self-testing functions.

Proposition 1 (Minimizer of the self-test loss function.) *Let u_l be a weak solution to $R_{\phi_*}[u_l] = f_l$ for each $1 \leq l \leq L$. The self-test loss function in (2.2) with $C_0 =$*

$\sum_{l=1}^L \langle R_{\phi_*}[u_l], v_{\phi_*}[u_l] \rangle$ can be written as

$$\mathcal{E}_{\mathcal{D}}(\phi) = \sum_{l=1}^L \langle R_{\phi-\phi_*}[u_l], v_{\phi-\phi_*}[u_l] \rangle \quad (2.3)$$

and it has ϕ_* as a minimizer. In particular, ϕ_* is the unique minimizer in a linear space $\mathcal{H}$ if and only if there exists $l \in \{1, \dots, L\}$ such that $\langle R_{\phi}[u_l], v_{\phi}[u_l] \rangle > 0$ for every nonzero $\phi \in \mathcal{H}$. Also, any minimizer ϕ_0 of the self-test loss function is a solution to the equation

$$\sum_{l=1}^L \langle R_{\phi_0}[u_l] - f_l, v_{\psi}[u_l] \rangle = 0, \quad (2.4)$$

for all ψ such that $\sum_{l=1}^L \langle R_{\psi}[u_l], v_{\psi}[u_l] \rangle < \infty$.

Proof Given the above C_0 , Eq.(2.3) follows from

$$\begin{aligned} \mathcal{E}_{\mathcal{D}}(\phi) &= \sum_{l=1}^L [\langle R_{\phi}[u_l], v_{\phi}[u_l] \rangle - 2\langle f_l, v_{\phi}[u_l] \rangle + \langle R_{\phi_*}[u_l], v_{\phi_*}[u_l] \rangle] \\ &= \sum_{l=1}^L \langle R_{\phi-\phi_*}[u_l], v_{\phi-\phi_*}[u_l] \rangle, \end{aligned}$$

where the last equality follows from the facts that $R_{\phi}[u]$ and $v_{\phi}[u]$ are linear in ϕ , and that $\langle R_{\phi}[u_l], v_{\phi_*}[u_l] \rangle = \langle R_{\phi_*}[u_l], v_{\phi}[u_l] \rangle = \langle f_l, v_{\phi}[u_l] \rangle$. Then, ϕ_* is a minimizer by the positivity property. Also, this equation implies that the uniqueness of the minimizer in the linear space $\mathcal{H}$ is equivalent to the strict positivity of $\frac{1}{L} \sum_{l=1}^L \langle R_{\phi}[u_l], v_{\phi}[u_l] \rangle$ for every nonzero $\phi \in \mathcal{H}$. Thus, ϕ_* is the unique minimizer in $\mathcal{H}$ iff there exists $l \in \{1, \dots, L\}$ such that $\langle R_{\phi}[u_l], v_{\phi}[u_l] \rangle > 0$ for every nonzero $\phi \in \mathcal{H}$.

Lastly, since ϕ_0 is a minimizer of the loss function, we have, for any ψ s.t. $\mathcal{E}_{\mathcal{D}}(\phi_0 + \epsilon\psi) < \infty$,

$$0 = \frac{d}{d\epsilon} \mathcal{E}_{\mathcal{D}}(\phi_0 + \epsilon\psi) = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{E}_{\mathcal{D}}(\phi_0 + \epsilon\psi) - \mathcal{E}_{\mathcal{D}}(\phi_0)}{\epsilon} = \sum_{l=1}^L \langle R_{\phi_0}[u_l] - f_l, v_{\psi}[u_l] \rangle,$$

and it gives Eq.(2.4). $\square$

2.2 Example: Wasserstein Gradient Flow

We consider first the estimation of function-valued parameters in the Wasserstein gradient flow in (1.5) from data in Example 1. Here the diffusion constant can be either $\nu > 0$ or $\nu = 0$, and the diffusion rate function $h : \mathbb{R} \rightarrow \mathbb{R}$ satisfies that $r \mapsto r^d h(r^{-d})$ is convex non-increasing. Examples of such h include

$$h(s) = s \frac{1}{m-1} s^{m-1} = \begin{cases} \frac{1}{m-1} s^m, & m > 1, \\ s \log s, & m = 1, \end{cases} \quad (2.5)$$

where we use the convention $\frac{1}{m-1} \rho^{m-1} = \log \rho$ when $m = 1$. In particular, when $m = 1$, we have $h'(u) = 1 + \log u$ and $\nabla \cdot (u \nabla h'(u)) = \nabla \cdot [u \nabla (1 + \log u)] = \Delta u$, and (1.5) becomes

$$\partial_t u = \nu \Delta u + \nabla \cdot (u \nabla [V + \Phi * u]), \quad x \in \mathbb{R}^d, t > 0. \quad (2.6)$$

This is the mean-field equation for the large N limit of the interacting particle system,

$$dX_t^i = -[\nabla V(X_t^i) + \frac{1}{N} \sum_{j=1}^N \nabla \Phi(X_t^i - X_t^j)]dt + \sqrt{2\nu d} W_t^i, \quad 1 \leq i \leq N, \quad (2.7)$$

where $(W_t^i)_{1 \leq i \leq N}$ are $\mathbb{R}^d$ -valued independent Brownian motions, and $(X_0^i)_{1 \leq i \leq N}$ are independent samples of distribution $u(\cdot, 0)$; see e.g., [19, 20].

Self-test loss function for estimating (h, Φ, V) . The task is to estimate the parameter $\phi = (h, \Phi, V)$ in the operator $R_\phi[u]$ in (1.5). Its self-testing function is

$$v_\phi[u] := \nu h'(u) + \Phi * u + V. \quad (2.8)$$

It is direct to verify the self-testing properties in (2.1): clearly, the symmetry and linearity hold; the positivity holds since by integration by parts, $\langle R_\phi[u], v_\phi[u] \rangle = \int_{\mathbb{R}^d} u |\nabla[\nu h'(u) + \Phi * u + V]|^2 dx \geq 0$, for all ϕ such that $\langle R_\phi[u], v_\phi[u] \rangle$ is well-defined.

Hence, the self-test loss function for data $(u(t, x) : t \in [0, T], x \in \mathbb{R}^d)$ is (1.7). Its minimizer matches the energy dissipation of the gradient flow, which we explore in Section 3.

Self-test loss function for estimating (Φ, V) . Assume that $\Phi(-x) = \Phi(x)$. Consider the problem of estimating (Φ, V) in the mean-field equation (2.6), i.e., estimating the parameter $\phi = (\Phi, V)$ in the (weak-form) operator $R_\phi[u] = -\nabla \cdot [u \nabla (\Phi * u + V)]$. The self-testing function is $v_\phi[u] = \Phi * u + V$, and $\langle R_\phi[u], v_\phi[u] \rangle = \int_{\mathbb{R}^d} u |\nabla \Phi * u + \nabla V|^2 dx$. Thus, the self-test loss function is

$$\begin{aligned} \mathcal{E}_u(\Phi, V) &= \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} [u |\nabla \Phi * u + \nabla V|^2 - 2(\partial_t u - \nu \Delta u)(\Phi * u + V)] dx dt \\ &= \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} [u |\nabla \Phi * u + \nabla V|^2 + 2\nu u (\Delta \Phi * u + \Delta V)] dx dt \\ &\quad - \frac{2}{T} \int_{\mathbb{R}^d} [u(T, x) [\Phi * u(T, x)/2 + V(x)] - u(0, x) [\Phi * u(0, x)/2 + V(x)]] dx, \end{aligned} \quad (2.9)$$

where the last equality follows from integration by parts and $\Phi(-x) = \Phi(x)$.

In practice, when the data is discrete, as in (1.6), we approximate the integrals in (2.9) using numerical methods, such as Riemann sums.

2.3 Example: Elliptic Diffusion Operators

To estimate $a : \mathbb{R}^d \rightarrow \mathbb{R}$ in Example 2, we have $R_a[u] = -\Delta(au) : C_c^1(\mathbb{R}^d) \rightarrow \mathbb{Y}$. Here $\mathbb{Y}$ is a Banach space such that $BV^* \subset \mathbb{Y}$ and $\mathbb{Y}^* \subset BV$, where BV denotes the space of functions with bounded variation. The self-testing function is $v_a[u] = au \in C_c^1(\mathbb{R}^d)$, whose self-testing properties follow directly, in particular, $\langle R_a[u], v_a[u] \rangle = -\int_{\mathbb{R}^d} \Delta(au) au dx = \int_{\mathbb{R}^d} |\nabla(au)|^2 dx \geq 0$ for all $a \in C_c^1(\mathbb{R}^d)$. Hence, the self-test loss function for a single data pair (u, f) is

$$\mathcal{E}_{(u, f)}(a) = \langle R_a[u] - 2f, v_a[u] \rangle = \int_{\mathbb{R}^d} [|\nabla(au)|^2 - 2f au] dx. \quad (2.10)$$

Approximating the integrals by Riemann sums with the data in (1.9), we obtain an empirical self-test loss function

$$\mathcal{E}_{\mathcal{D}_2}(a) = \frac{1}{L} \sum_{l=1}^L \mathcal{E}_{(u_l, f_l)}(a) = \frac{1}{NL} \sum_{i,l=1}^{N,L} [|\nabla(au_l)(x_i)|^2 - 2f_l(x_i)a(x_i)u_l(x_i)] |\Delta x_i|.$$

2.4 Example: Sequential Ensembles of Unlabeled Data

To estimate the potentials from sequential ensembles of unlabeled data $(X_{t_l}^{i_l, (m)}, 1 \leq i_l \leq N)$ in Example 3, we consider the empirical measures of the data

$$\mu_N^{(m)}(x, t_l) = \frac{1}{N} \sum_{i_l=1}^N \delta_{X_{t_l}^{i_l, (m)}}(x).$$

We construct a self-test loss function using the fact that the empirical measure $\mu_N(x, t) := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}(x)$ with $(X_t^i, 1 \leq i \leq N)$ satisfying (1.10) is a weak solution to equation

$$\partial_t \mu_N = \nabla \cdot (\mu_N \nabla [V + \Phi * \mu_N]), \quad \mu_N(\cdot, t) \in \mathcal{P}_2(\mathbb{R}^d), \quad t > 0. \quad (2.11)$$

In other words, for any function $v \in C^2(\mathbb{R}^d)$,

$$\langle \partial_t \mu_N, v \rangle = \langle \nabla \cdot (\mu_N (\nabla \Phi * \mu_N + V)), v \rangle = - \langle \mu_N (\nabla \Phi * \mu_N + V), \nabla v \rangle,$$

where the second equality follows from integration by parts. In fact, the above equation holds by the chain rule with the differential equation (1.10):

$$\begin{aligned} \langle \partial_t \mu_N, v \rangle &= \frac{1}{N} \sum_{i=1}^N \frac{d}{dt} v(X_t^i) = \frac{1}{N} \sum_{i=1}^N \frac{dX_t^i}{dt} \cdot \nabla v(X_t^i) \\ &= -\frac{1}{N} \sum_{i=1}^N \left(\frac{1}{N} \sum_{j=1}^N \nabla \Phi(X_t^i - X_t^j) + \nabla V(X_t^i) \right) \cdot \nabla v(X_t^i) \\ &= - \langle \mu_N (\nabla \Phi * \mu_N + V), \nabla v \rangle \end{aligned}$$

and by noticing that $\nabla \Phi * \mu_N(x) = \frac{1}{N} \sum_{j=1}^N \nabla \Phi(x - X_t^j)$.

Thus, we consider the weak-form operator $R_\Phi[u] = -\nabla \cdot [u \nabla (\Phi * u + V)]$ with output $f = \partial_t u$. The self-testing function is $v_\Phi[u] = \Phi * u + V$, and $\langle R_\Phi[u], v_\Phi[u] \rangle = \int_{\mathbb{R}^d} u |\nabla \Phi * u + \nabla V|^2 dx$. Then, the self-test loss function is (2.9) with $v = 0$. Using the data-induced empirical measures $\{\mu_N^{(m)}(\cdot, t_l)\}$, we have a self-test loss function

$$\begin{aligned} \mathcal{E}_{\mathcal{D}_3}(\Phi, V) &= \frac{1}{LMN} \sum_{l,i,m=1}^{L,N,M} \left| \frac{1}{N} \sum_{j=1}^N \nabla \Phi(X_{t_l}^{i, (m)} - X_{t_l}^{j, (m)}) + \nabla V(X_{t_l}^{i, (m)}) \right|^2 dt \\ &\quad - \frac{2}{LMN} \sum_{i,m=1}^{N,M} \left[\frac{1}{N} \sum_{j=1}^N \Phi(X_{t_l}^{i, (m)} - X_{t_l}^{j, (m)}) + V(X_{t_l}^{i, (m)}) \right] \Big|_{t_1}^{t_L}. \quad (2.12) \end{aligned}$$

Note that this empirical loss function does not use the trajectory information of any single particle, and it uses exactly the ensemble data of unlabeled particles. We demonstrate the application of this loss function in Section 5.3.

Remark 1 Eq.(2.11) is the same as (1.5) with $v = 0$ and the empirical measures $(\mu_N(\cdot, t), t \geq 0)$ form a Wasserstein gradient flow on $\mathcal{P}_2(\mathbb{R}^d)$. However, it is not the Liouville equation of the ODE in (1.10), since the Liouville equation governs the evolution of the joint distribution on $\mathbb{R}^{Nd}$. Similarly, the mean-field equation in (2.6) is not the Fokker-Planck equation of the SDE in (2.7), but we can use it to derive the same self-test loss function for the SDE with sequential ensembles of unlabeled data $\mathcal{D}_3$.

3 Connection with Energy Conservation and Likelihood

This section connects the self-test loss to two fundamental principles: the energy conservation law of gradient flows and the maximal likelihood principle for inference in stochastic differential equations (SDEs). We show that the self-test loss is designed to match the energy dissipation of a gradient flow, and that its minimizer satisfies the corresponding *energy conservation law* for the observed data. Moreover, the first variation of the free energy naturally yields a self-testing function. These results are illustrated through the Wasserstein and parabolic gradient flow examples. Finally, we show that, for SDEs, the self-test loss coincides with the expected negative log-likelihood ratio.

3.1 Matching Energy Dissipation for Gradient Flow

We first define the self-test loss function for a generic gradient flow whose free energy depends linearly on the parameter.

Consider the estimation of the function-valued parameter ϕ in the free energy $E_\phi : \mathbb{M} \rightarrow \mathbb{R}$, where $\mathbb{M}$ is a metric space, from a gradient flow path $u_{[0, T]} := (u(t, \cdot), t \in [0, T]) \subset \mathbb{M}$. Here, the gradient flow satisfies the equation

$$\partial_t u = -A_u \frac{\delta E_\phi}{\delta u}, \quad (3.1)$$

where $\partial_t u \in T_u \mathbb{M}$, $A_u : T_u^* \mathbb{M} \rightarrow T_u \mathbb{M}$ is a nonnegative definite operator from the cotangent plane $T_u^* \mathbb{M}$ to the tangent plane $T_u \mathbb{M}$, and $\frac{\delta E_\phi}{\delta u} \in T_u^* \mathbb{M}$ is the Fréchet derivative (also called the first variation) of the free energy. Its weak form reads

$$\langle \partial_t u, g \rangle + \left\langle A_u \frac{\delta E_\phi}{\delta u}, g \right\rangle = 0, \quad \forall g \in T_u^* \mathbb{M},$$

where $\langle \cdot, \cdot \rangle$ is the dual pair on $T_u \mathbb{M} \times T_u^* \mathbb{M}$.

We define a self-test loss function for estimating ϕ by connecting the gradient flow with the weak form operator R_ϕ in (1.1) and its self-testing function $v_\phi[u]$ as follows:

$$R_\phi[u] = A_u \frac{\delta E_\phi}{\delta u}, \quad v_\phi[u] = \frac{\delta E_\phi}{\delta u}. \quad (3.2)$$

The following assumptions on the gradient flow ensure the self-testing properties in (2.1).

Assumption 3.1 Assume the gradient flow in (3.1) satisfies the following properties.

(i) The operator A_u is linear, nonnegative definite, and symmetric: $\forall \xi, \eta \in T_u^* \mathbb{M}$,

$$\begin{aligned} \text{linear:} & \quad A_u(\xi + \eta) = A_u \xi + A_u \eta; \\ \text{symmetric:} & \quad \langle A_u \xi, \eta \rangle = \langle \xi, A_u \eta \rangle; \\ \text{nonnegative definite:} & \quad \langle A_u \xi, \xi \rangle \geq 0. \end{aligned} \quad (3.3)$$

Here $\langle \cdot, \cdot \rangle$ are dual pair on $T_u \mathbb{M} \times T_u^* \mathbb{M}$.

- (ii) The free energy E_ϕ depends on ϕ linearly. Consequently, $\frac{\delta E_\phi}{\delta u}$ is also linear in ϕ , i.e., $\frac{\delta E_{\phi+\psi}}{\delta u} = \frac{\delta E_\phi}{\delta u} + \frac{\delta E_\psi}{\delta u}$ for all ϕ, ψ such that the energy function is well-defined.

Definition 2 (Self-test loss function for gradient flow) Consider the problem of estimating ϕ in the gradient flow (3.1) satisfying Assumption 3.1.

Given continuous time data $u_{[0,T]} := (u(t, \cdot), t \in [0, T])$, a self-test loss function is

$$\mathcal{E}_{u_{[0,T]}}(\phi) = 2[E_\phi(u(T, \cdot)) - E_\phi(u(0, \cdot))] + \int_0^T \left\langle A_u \frac{\delta E_\phi}{\delta u}, \frac{\delta E_\phi}{\delta u} \right\rangle dt. \quad (3.4)$$

The next theorem shows that the self-test loss function has the true parameter ϕ_* as a minimizer, and that its minimizer satisfies energy conservation for the data flow. We postpone its proof to Appendix A.1.

Theorem 3.2 (Minimizer of the loss function) *The minimizer of the loss function $\mathcal{E}_{u_{[0,T]}}(\phi)$ in (3.4) of Definition 2 satisfies the following properties.*

- (a) *The true parameter ϕ_* is a minimizer and $\mathcal{E}_{u_{[0,T]}}(\phi_*) = -\int_0^T \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi_*}}{\delta u} \right\rangle dt$.*
 (b) **Uniqueness.** *The minimizer is unique in a linear parameter space $\mathcal{H}$ if*

$$\int_0^T \left\langle A_u \frac{\delta E_\phi}{\delta u}, \frac{\delta E_\phi}{\delta u} \right\rangle dt > 0, \quad \forall \phi \in \mathcal{H}, \phi \neq 0. \quad (3.5)$$

- (c) **Energy conservation.** *A minimizer ϕ_0 of $\mathcal{E}_{u_{[0,T]}}(\phi)$ satisfies the energy conservation for the data $u_{[0,T]}$. That is, the energy change $E_{\phi_0}[u(T, \cdot)] - E_{\phi_0}[u(0, \cdot)]$ matches the total energy dissipation $-\int_0^T \left\langle A_u \frac{\delta E_{\phi_0}}{\delta u}, \frac{\delta E_{\phi_0}}{\delta u} \right\rangle dt$ along the flow $u_{[0,T]}$:*

$$E_{\phi_0}[u(T, \cdot)] - E_{\phi_0}[u(0, \cdot)] = -\int_0^T \left\langle A_u \frac{\delta E_{\phi_0}}{\delta u}, \frac{\delta E_{\phi_0}}{\delta u} \right\rangle dt. \quad (3.6)$$

Example: the Wasserstein gradient flow. We show first that the Wasserstein gradient flow in Eq.(1.5) satisfies Assumption 3.1; thus, its self-test loss function in (2.2) aims to match the energy dissipation in the data.

Let $\mathbb{M} := (\mathcal{P}_2(\mathbb{R}^d), W_2)$ be the space of probability measures with finite second moments endowed with the Wasserstein-2 metric W_2 . Recall that for any convex functional $E(u)$ over $\mathbb{M}$, the gradient is $\nabla^{W_2} E(u) = -\nabla \cdot (u \nabla \frac{\delta E}{\delta u})$, where ∇ is the gradient with respect to x (see e.g., [3, 44]). Then, a gradient flow in $\mathbb{M}$ is

$$\partial_t u = -\nabla^{W_2} E = \nabla \cdot \left(u \nabla \frac{\delta E}{\delta u} \right) = -A_u \frac{\delta E}{\delta u} \quad \text{with } A_u \xi := -\nabla \cdot (u \nabla \xi). \quad (3.7)$$

Clearly, the operator $A_u : T_u^* \mathbb{M} \rightarrow T_u \mathbb{M}$ is linear, non-negative definite and symmetric, i.e., it satisfies Assumption 3.1(i).

To connect with Eq.(1.5), consider the free energy with parameter $\phi = (h, \Phi, V)$:

$$E_\phi(u) = v \int h(u) + \frac{1}{2} \iint \Phi(x-y) u(x) u(y) dx dy + \int V(x) u(x) dx.$$

Here, the first term is called entropy (named when $h(s) = s \log s$) or internal energy in general, and the second and third terms are called interaction energy and potential energy. Since $\Phi(x) = \Phi(-x)$, the Fréchet derivative of this energy function is

$$\frac{\delta E_\phi}{\delta u} = v h'(u) + \Phi * u + V. \quad (3.8)$$

Then, the W_2 -gradient flow equation (3.7) becomes Eq.(1.5).

In particular, note that both E_ϕ and its derivative $\frac{\delta E_\phi}{\delta u}$ in (3.8) are linear in ϕ . In other words, Assumption 3.1(ii) holds. Thus, we can define the self-test loss function in (3.4). Meanwhile, note that the above $\frac{\delta E_\phi}{\delta u}$ is exactly the self-testing function $v_\phi[u]$ in (2.8). Thus, this self-test loss function agrees with the one in (1.7).

Thus, by Theorem 3.2, the self-test loss function has ϕ_* as a minimizer, and any of its minimizers matches the energy conservation for the data flow.

Example: the parabolic gradient flow. Consider next estimating the coefficient $a(x)$ from data $u_{[0,T]}$ of the parabolic (or H^{-1}) gradient flow

$$\partial_t u = \Delta(a(x)u), \quad x \in \mathbb{T}^d, \quad (3.9)$$

where $\mathbb{T}^d$ is the d -dimensional torus. It is a H^{-1} gradient flow of the free energy $E_a(u) := \frac{1}{2} \int a(x)u^2 dx$ since $\nabla^{H^{-1}} E = -\Delta \frac{\delta E_a}{\delta u}$ and $\frac{\delta E_a}{\delta u} = au$. In other words, Eq.(3.9) can be written as

$$\partial_t u = -\nabla^{H^{-1}} E = -A_u \frac{\delta E_a}{\delta u} \quad \text{with } A_u \xi = -\Delta \xi.$$

Clearly, Assumption 3.1 holds since (i) the $A_u : H^1 \rightarrow H^{-1}$ is linear, nonnegative definite and symmetric, and (ii) the energy function E_a and its derivative $\frac{\delta E_a}{\delta u}$ are linear in a . Thus, by (3.4) with integration by parts, the self-test loss function is

$$\mathcal{E}_{u_{[0,T]}}(a) = \int_{\mathbb{T}^d} [u(T, x)^2 - u(0, x)^2] a(x) dx + \int_0^T \int_{\mathbb{T}^d} |\nabla(au)|^2 dx dt.$$

It is the time-integrated version of the loss function (2.10) with $f = \partial_t u$ for Example 2.

3.2 Expected Likelihood Ratio of the McKean-Vlasov SDE

Next, we show that for the McKean-Vlasov SDE, the self-test loss function of its Fokker-Planck equation coincides with the expectation of the negative log-likelihood ratio (see Appendix A.1 for its proof).

Theorem 3.3 *Consider the problem of estimating the potentials $V_*, \Phi_* : \mathbb{R}^d \rightarrow \mathbb{R}$ in the McKean-Vlasov SDE*

$$\begin{cases} d\bar{X}_t = -\nabla[V_*(\bar{X}_t) + \Phi_* * u(\bar{X}_t, t)]dt + \sqrt{2\nu}dB_t, \\ u(x, t) = \mathbb{E}[\delta_{\bar{X}_t}(x)]. \end{cases} \quad (3.10)$$

Suppose that the data is $u_{[0,T]} := (u(t, x), t \in [0, T], x \in \mathbb{R}^d)$, where $u(t, \cdot)$ is the probability distribution of $\bar{X}_t$. Then, the self-test loss function in (2.9) for the weak form Fokker-Planck equation in (2.6) is the expectation of the negative log-likelihood ratio $\mathcal{E}_{\bar{X}_{[0,T]}}(\Phi, V)$ of the path $\bar{X}_{[0,T]}$, i.e., $\mathcal{E}_{u_{[0,T]}}(\Phi, V) = \frac{2\nu}{T} \mathbb{E}[\mathcal{E}_{\bar{X}_{[0,T]}}(\Phi, V)]$.

A key advantage of the self-test loss function is its applicability for both $\nu > 0$ and $\nu = 0$. In contrast, the likelihood-based approach requires $\nu > 0$, as this condition is essential for applying the Girsanov theorem to define a non-degenerate measure on the path space. The self-test loss function, however, imposes no such constraint on ν . Notably, when $\nu = 0$, the SDE reduces to an ordinary differential equation (ODE). When the ODE has a random initial condition, the self-test loss function is derived from the Liouville equation governing the distribution flow.

Importantly, as the next proposition shows, we can write the self-test loss function as a combination of expectations for probability flows. This allows Monte Carlo approximation of the loss function, which is particularly useful for high-dimensional problems when the data consists of sequential ensembles of samples.

Corollary 1 *The loss function of (3.10) with $v \geq 0$ can be written as expectations:*

$$\begin{aligned} \mathcal{E}_{u_{[0,T]}}(\Phi, V) = & \frac{1}{T} \int_0^T \left(\mathbb{E} \left| \mathbb{E}[\nabla \Phi(Z_t) | \bar{X}_t] + \nabla V(\bar{X}_t) \right|^2 + 2v \mathbb{E}[\Delta \Phi(Z_t) + \Delta V(\bar{X}_t)] \right) dt \\ & - 2(\mathbb{E}[\Phi(Z_T) + V(\bar{X}_T)] - \mathbb{E}[\Phi(Z_0) + V(\bar{X}_0)]), \end{aligned} \quad (3.11)$$

where $Z_t = \bar{X}_t - \bar{X}'_t$ and $\bar{X}'_t$ is an independent copy of $\bar{X}_t$.

Proof Recall that $u(\cdot, t)$ is the probability density function of $\bar{X}_t$. Hence, we can write the integrals as expectations, for example, $\int_{\mathbb{R}^d} u |\nabla V|^2 dx = \mathbb{E}[|\nabla V(\bar{X}_t)|^2]$. In particular, note that $\Phi * u(\bar{X}_t) = \mathbb{E}[\Phi(\bar{X}_t - \bar{X}'_t) | \bar{X}_t]$, where $\bar{X}'_t$ is an independent copy of $\bar{X}_t$. Then, we have $\int_0^T \int_{\mathbb{R}^d} u |\nabla \Phi * u + \nabla V|^2 dx dt = \int_0^T \mathbb{E} \left| \mathbb{E}[\nabla \Phi(Z_t) | \bar{X}_t] + \nabla V(\bar{X}_t) \right|^2 dt$.

Meanwhile, note that $\mathbb{E}[\Phi * u(\bar{X}_t)] = \mathbb{E}[\mathbb{E}[\Phi(\bar{X}_t - \bar{X}'_t) | \bar{X}_t]] = \mathbb{E}[\Phi(Z_t)]$. Then, with integration by parts, we can write

$$\int_{\mathbb{R}^d} (\Phi * u + V)(\partial_t u - v \Delta u) dx = \partial_t \mathbb{E}[\Phi(Z_t) + V(\bar{X}_t)] - v \mathbb{E}[\Delta \Phi(Z_t) + \Delta V(\bar{X}_t)].$$

Integrate in time over $[0, T]$, we obtain (3.11). $\square$

4 Identifiability and Well-Posedness

The quadratic structure of the self-test loss functions provides a framework for analyzing the identifiability of the (function-valued) parameters and the well-posedness of the inverse problems. Notably, since no prior information is assumed for the unknown parameters, we define *adaptive spaces* that depend on both the operator and the data. These spaces capture the limited information available for parameter estimation and provide the appropriate setting for studying the identifiability and well-posedness of the inverse problems.

We demonstrate the approach by estimating h , Φ , and V in the operator defined in (1.5):

$$R_\phi[u] := R_{(h, \Phi, V)}[u] = -\nabla \cdot (u \nabla [v h'(u) + \Phi * u + V]) = f.$$

We start by estimating each parameter individually, the other two are known, in Sections 4.1 and 4.3. Finally, we address the joint estimation of all three parameters. Notably, we establish that the inverse problems for estimating h and V are well-posed, while the estimation of Φ is ill-posed due to the nonlocal nature of the interaction.

Throughout this section, we construct the parameter spaces using continuum data of input-output pairs $(u_l, f_l)_{l=1}^L$. In practice, discrete data approximates continuum data under appropriate smoothness conditions, as specified in the following assumption.

Assumption 4.1 The data $\{(u_l, f_l)\}_{l=1}^L$ satisfies $f_l = R_{\phi_*}[u_l]$, where $\phi_* = (h_*, \Phi_*, V_*)$, and $\{u_l\}_{l=1}^L \subset \mathbb{X} := C_c^1(\mathbb{R}^d)$, i.e., each u_l has continuous derivatives and compact support.

Generalization to non-smooth data u_l is possible in specific settings. For instance, in the absence of the diffusion term (e.g., $v = 0$), it suffices for each u_l to be a continuous probability density function supported on a compact subset of $\mathbb{R}^d$ for the results in Sections 4.2–4.3.

4.1 Estimating the Diffusion Rate: Well-Posed

Consider first estimating the diffusion rate function $h : \mathbb{R}^+ \rightarrow \mathbb{R}$ when Φ and V are given. We rewrite the equation $R_\phi[u] = f$ to isolate the unknown:

$$R_h[u] := -\nabla \cdot [u \nabla (v h'(u))] = -\nabla \cdot [v h''(u) u \nabla u] = f + \nabla \cdot (u \nabla [\Phi * u + V]) := \tilde{f}. \quad (4.1)$$

Evidently, only h'' is identifiable, since R_h depends on h solely through h'' . Accordingly, we formulate the self-test loss directly in terms of h'' .

Using the same arguments in Sect. 2.2, one can verify that $v_h[u] = h'(u)$ is a self-testing function, and the self-test loss function is

$$\mathcal{E}_1(h'') = \sum_{l=1}^L \langle R_h[u_l] - 2\tilde{f}_l, v_h[u_l] \rangle + C_0 \quad (4.2)$$

with C_0 being an arbitrary constant. Here we used the notation $\mathcal{E}_1$ to indicate that this loss function is for estimating h .

Given data $\{u_l\}$ with a compact support, we take the parameter space for h'' to be $L^2_{\rho_1}$, where the measure ρ_1 is defined through its density function

$$\dot{\rho}_1(r) = \sum_{l=1}^L \int_{\mathbb{R}^d} \delta(u_l(x) - r) |\nabla u_l(x)|^2 u_l(x) dx \quad (4.3)$$

with $\delta(\cdot)$ being the Dirac delta function. For any h'' in this space, the quadratic term in the loss function is well-defined since

$$\sum_{l=1}^L \langle R_h[u_l], v_h[u_l] \rangle = \sum_{l=1}^L \int_{\mathbb{R}^d} u_l(x) |\nabla u_l(x)|^2 h''(u_l(x))^2 dx = \int_{\mathbb{R}^+} h''(r)^2 \dot{\rho}_1(r) dr,$$

where we used the fact that

$$\int_{\mathbb{R}^d} u(x) |\nabla u(x)|^2 h''(u(x))^2 dx = \int_{\mathbb{R}^+} h''(r)^2 \int_{\mathbb{R}^d} u(x) |\nabla u(x)|^2 \delta(u(x) - r) dx dr.$$

The next proposition presents the well-posedness of estimating h'' in $L^2_{\rho_1}$.

Proposition 2 *Given data $\{(u_l, f_l)\}_{l=1}^L$ satisfying Assumption 4.1, the self-test loss function in (4.2) for estimating h'' in Eq.(4.1) has a unique minimizer in $L^2_{\rho_1}$ with ρ_1 defined in (4.3). In particular, the inverse problem of estimating h'' is well-posed.*

We postpone its proof, along with the proofs of the remaining propositions in this section, to Appendix A.2. Since this inverse problem is well-posed, regularization in practice (e.g., Section 5.1) serves primarily to smooth the estimator or to filter errors from noise and discretization.

4.2 Estimating the Kinetic Potential: Well-Posed

Similarly, we next estimate the potential $V : \mathbb{R}^d \rightarrow \mathbb{R}$ assuming that h and Φ are given. Rewriting $R_\phi[u] = f$ to isolate V gives

$$R_V[u] := -\nabla \cdot (u \nabla V) = f + \nabla \cdot (u \nabla [\Phi * u + v h'(u)]) =: \tilde{f}. \quad (4.4)$$

Since R_V depends only on ∇V , we can identify V only up to an additive constant. Accordingly, we formulate the self-test loss directly in terms of ∇V . With $v_V[u] = V$ as a self-testing function, the self-test loss function is

$$\mathcal{E}_2(\nabla V) = \sum_{l=1}^L \langle R_V[u_l] - 2\tilde{f}_l, V \rangle = \|\nabla V\|_{L^2_{\rho_2}}^2 - 2 \left\langle \sum_{l=1}^L \tilde{f}_l, V \right\rangle, \quad (4.5)$$

where we got $\sum_{l=1}^L \langle R_V[u_l], V \rangle = \sum_{l=1}^L \int_{\mathbb{R}^d} u_l(x) |\nabla V(x)|^2 dx = \|\nabla V\|_{L^2_{\rho_2}}^2$ by integration by parts, and the data-dependent measure ρ_2 is defined by its density function

$$\dot{\rho}_2(x) = \sum_{l=1}^L u_l(x). \quad (4.6)$$

The next proposition shows that the inverse problem of estimating $\nabla V \in L^2_{\rho_2}(\mathbb{R}^d, \mathbb{R}^d)$ is well-posed (see its proof in Appendix A.2). For estimating V , the inverse problem is well-posed in $\mathcal{H}_0 := \{g \in H^1_{\rho_2}(\mathbb{R}^d; \mathbb{R}); \int_{\mathbb{R}^d} g \rho_2 dx = 0\}$ when the measure ρ_2 satisfies the Poincaré inequality. Here, $H^1_{\rho_2}(\mathbb{R}^d; \mathbb{R}) := \{g \in L^2_{\rho_2} : |\nabla g| \in L^2_{\rho_2}\}$.

Proposition 3 *Consider the problem of estimating ∇V or V in Eq.(4.4) from data $\{(u_l, \tilde{f}_l)\}_{l=1}^L$ satisfying Assumption 4.1. Let ρ_2 be the measure defined in (4.6).*

- *For estimating ∇V , the self-test loss function in (4.5) is uniformly convex and has a unique minimizer in $L^2_{\rho_2}(\mathbb{R}^d, \mathbb{R}^d)$. Consequently, the inverse problem is well-posed.*
- *For estimating V , assume that ρ_2 satisfies the Poincaré inequality, i.e.,*

$$\int_{\mathbb{R}^d} |g|^2 \rho_2 dx \leq c \int_{\mathbb{R}^d} |\nabla g|^2 \rho_2 dx, \quad \forall g \in H^1_{\rho_2} \text{ with } \int_{\mathbb{R}^d} g \rho_2 dx = 0. \quad (4.7)$$

Then the self-test loss function in (4.5), when viewed as a functional of V in the space $\mathcal{H}_0 := \{g \in H^1_{\rho_2}(\mathbb{R}^d; \mathbb{R}); \int_{\mathbb{R}^d} g \rho_2 dx = 0\}$ is uniformly convex and has a unique minimizer $\hat{V}$ satisfying

$$-\nabla \cdot (\rho_2(\nabla \hat{V})) = \sum_{l=1}^L \tilde{f}_l. \quad (4.8)$$

We remark that the assumption of ρ_2 satisfying the Poincaré inequality in (4.7) is mild, and it is equivalent to the spectral gap condition on ρ_2 when ρ_2 is a probability measure; see, e.g., [1]. For instance, $\rho_2(x) = e^{-W(x)}$ with $W \in C^2(\mathbb{R}^d)$ and $\nabla^2 W(x) \geq \frac{1}{c} I_d$ for all x , or ρ_2 is supported on a bounded connected domain, bounded above, and bounded below away from 0. When ρ_2 satisfies this assumption, the potential V can be uniquely recovered up to a constant in $H^1_{\rho_2}$ since it is identifiable in $\mathcal{H}_0$. However, the minimizer is nonunique without this assumption or beyond $H^1_{\rho_2}$, as shown in the next one-dimensional example. Assume that $d = 1$ and $\rho_2(x) = e^{-x}$. Then, if V is a solution to (4.8), so is $V + e^x$ since $\nabla \cdot (\rho_2(\nabla e^x)) = 0$.

4.3 Estimating the Interaction Potential: Ill-Posed

The inverse problem of estimating $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ differs fundamentally from the previous two, as it is ill-posed due to its deconvolution structure. Here, we estimate Φ in

$$R_\Phi[u] := -\nabla \cdot (u \nabla (\Phi * u)) = f + \nabla \cdot (u \nabla [V + v h'(u)]) =: \tilde{f}, \quad (4.9)$$

where V and h' are given. As R_Φ depends on Φ only through $\nabla\Phi$, we can identify Φ only up to an additive constant.

Denote $F[u] = \Phi * u + V_* - V + \nu h'_*(u) - \nu h'(u)$ and note that $\tilde{f} = -\nabla \cdot (u \nabla[F[u]])$. With a self-testing function $v_\Phi[u] = \Phi * u$, the self-test loss function of $\nabla\Phi$ is

$$\begin{aligned} \mathcal{E}_3(\nabla\Phi) &= \frac{1}{2} \sum_{l=1}^L \langle R_\Phi[u_l] - 2\tilde{f}_l, \Phi * u_l \rangle \\ &= \frac{1}{2} \sum_{l=1}^L \int_{\mathbb{R}^d} u_l (|\nabla\Phi * u|^2 - 2\nabla F[u_l] \cdot \nabla\Phi * u_l) dx. \end{aligned} \quad (4.10)$$

The independent variable of Φ is the pairwise difference of particle positions, while the data u is the distribution of each particle. To quantify the exploration of the independent variable of Φ by data, we define a measure ρ_3 with a density function

$$\dot{\rho}_3(y) \propto \sum_{l=1}^L \int u_l(x) u_l(x - y) dx. \quad (4.11)$$

It extends the exploration measure defined in [23, 24] for radial interacting potentials.

Let $L_{\bar{G}} : L_{\rho_3}^2 \rightarrow L_{\rho_3}^2$ be an integral operator defined by

$$\begin{aligned} L_{\bar{G}} \nabla\Phi(y) &= \int \bar{G}(y, y') \nabla\Phi(y') \rho_3(dy') \quad \text{with } \bar{G}(y, y') = \frac{G(y, y')}{\dot{\rho}_3(y) \dot{\rho}_3(y')} \mathbf{1}_{\dot{\rho}_3(y) \dot{\rho}_3(y') > 0}, \\ G(y, y') &= \sum_{l=1}^L \int u_l(x) u_l(x - y) u_l(x - y') dx. \end{aligned} \quad (4.12)$$

Here $\bar{G}(y, y')$ is square integrable by Assumption 4.1; see [24].

The next proposition shows that we can only identify $\nabla\Phi \in \text{Null}(L_{\bar{G}})^\perp \subset L_{\rho_3}^2$, and the inverse problem of estimating $\nabla\Phi$ is ill-posed.

Proposition 4 *Consider the problem of estimating $\nabla\Phi$ in Eq.(4.9) from data $\{(u_l, f_l)\}_{l=1}^L$ satisfying Assumption 4.1. Let ρ_3 be the measure defined in (4.11). The Hessian of the quadratic self-test loss function in (4.10) is the compact operator $L_{\bar{G}}$ on $L_{\rho_3}^2(\mathbb{R}^d, \mathbb{R}^d)$ defined in (4.12). Consequently, the inverse problem of finding its minimizer in (7.3) is ill-posed.*

A regularization is necessary to obtain a stable solution for this ill-posed inverse problem of estimating $\nabla\Phi$. In particular, when $\text{Null}(L_{\bar{G}}) \neq \{0\}$, it is crucial to regularize only on $\text{Null}(L_{\bar{G}})^\perp$ in order to prevent the estimator from being contaminated by components in $\text{Null}(L_{\bar{G}})$. Data-adaptive RKHS regularization or priors, as proposed in [5, 26], employ the RKHS with reproducing kernel $\bar{G}$ and yield convergent estimators.

The ill-posedness in estimating $\nabla\Phi$ stems from the deconvolution structure of the problem. Consequently, even if additional properties are imposed on Φ , such as radial or symmetry, the inverse problem remains ill-posed. However, when the data u_l are contaminated by additive spatial noise, the operator $L_{\bar{G}}$ in (4.12) can become strictly positive definite, which in turn yields a well-posed inverse problem; see Section 5.2 for a numerical illustration.

4.4 Joint Estimation

Using the parameter spaces and operators in the previous sections, the joint estimation for $(h'', \nabla V, \nabla\Phi)$ takes place in the product space $L_{\rho_1}^2(\mathbb{R}^+) \otimes L_{\rho_2}^2(\mathbb{R}^d) \otimes L_{\rho_3}^2(\mathbb{R}^d)$. Meanwhile,

the self-test loss function can be written as

$$\mathcal{E}(h'', \nabla V, \nabla \Phi) := \sum_{l=1}^L \int_{\mathbb{R}^d} [u_l |\nabla[vh'(u_l) + \Phi * u_l + V]|^2 - 2f_l[vh'(u_l) + \Phi * u_l + V]] dx dt. \quad (4.13)$$

The next proposition shows the ill-posedness of estimating $(h'', \nabla V, \nabla \Phi)$.

Proposition 5 (Joint estimation) *Consider the problem of jointly estimating $h'', \nabla V, \nabla \Phi$ in Eq.(1.5) from data $\{(u_l, f_l)\}_{l=1}^L$ satisfying Assumption 4.1. Let ρ_1, ρ_2, ρ_3 be the measures defined in (4.3), (4.6) and (4.11), respectively. Then, the self-test loss function in (4.13) is not uniformly convex, and its Hessian (second variation) has a zero eigenvalue with eigenfunction $\phi = (0, \mathbf{c}, -\mathbf{c})$ for any nonzero $\mathbf{c} \in \mathbb{R}^d$. In particular, the joint estimation problem of finding the minimizer of the loss function is ill-posed.*

We remark that the zero eigenvalue of the loss function's Hessian roots in the fact that different pairs (Φ, V) and $(\Phi + \mathbf{c} \cdot x, V - \mathbf{c} \cdot x)$ produce the same value of the loss function, which has been noticed in [47]. To eliminate this degeneracy, we enforce symmetry on Φ , so that vectors of the form $(0, \mathbf{c}, -\mathbf{c})^T$ no longer lie in the admissible function space for Φ . In practice, this constraint can be implemented either by restricting to radial potentials (see Section 5.2) or by parametrizing Φ via $\Phi(x) = \tilde{\Phi}(x) + \tilde{\Phi}(-x)$, where $\tilde{\Phi}$ is a learnable function (e.g., a neural network as in Section 5.3).

5 Applications to Parametric and Nonparametric Estimations

We demonstrate applications of the self-test loss function in estimating the function parameters in the weak form operator $R_{(h, \Phi, V)}[u] = -\nabla \cdot (u \nabla[vh'(u) + \Phi * u + V])$ in (1.5) and its gradient flow. We consider parameter estimation for h in Section 5.1, nonparametric estimation for radial Φ in Section 5.2, and neural network regression for joint estimation of Φ and V in Section 5.3.

5.1 Parametric Estimation of the Diffusion Rate Function

Consider first a parametric estimation of h in the equation

$$R_h[u] := -\nabla \cdot (u[\nabla h'(u)]) = -\nabla \cdot [uh''(u)\nabla u] = f, \quad (5.1)$$

from data $\{(u_l(x_i), f_l(x_i))\}_{i=1, l=1}^{N, L}$, where $x_i \in [0, 1]$ is a uniform mesh and $u_l \in H_0^1((0, 1))$. Here, the diffusion rate function h is a power function in (2.5) with a parametric form

$$h_{\mathbf{c}}(s) = c_2 s^2 + c_3 \frac{1}{2} s^3 + c_4 \frac{1}{3} s^4 = \sum_{k=1}^{n_c} c_k e_k(s), \quad (5.2)$$

where $e_k(s) = \frac{1}{k-1} s^k$ for $k > 1$, and $n_c = 3$. Thus, the task is to estimate the parameters $\mathbf{c} = (c_2, c_3, c_4)$. Here we don't include the term $e_1(s) = s \log s$ because its second derivative $e_1''(s) = 1/s$ is singular at $s = 0$. Such a singularity leads to a singular function $e_1''(u_l(x))$ when $u_l(x)$ approaches zero at the boundaries, requiring additional numerical treatments when computing the loss function of h'' and the normal matrix for regression.

Synthetic Data generation. We generate data by adding noise to the values of analytically computed functions on the spatial mesh. Let the mesh points be $x_i = \{\frac{j}{N}, 1 \leq j \leq N\}$.

We obtain noisy data $\{u_l(x_i)\}$ by adding independent Gaussian noises $\mathcal{N}(0, \sigma^2/N)$ to the values of $u_l(x) = \sin(\pi l x)$ on the mesh for $l \in \{1, 2, 3\}$. Note that these functions are in $H_0^1((0, 1))$.

The data $\{f_l(x_i)\}$ are noisy observations of $R_{h_{\mathbf{c}^*}}[u_l](x)$ at the meshes:

$$f_l(x_i) = -R_{h_{\mathbf{c}^*}}[u_l](x_i) + \epsilon_{l,i} = -\sum_{k=2}^{n_c} c_k \nabla \cdot [u_l e_k''(u_l) \nabla u_l](x_i) + \epsilon_{l,i}$$

with parameter $\mathbf{c}^* = (c_2, c_3, c_4) = (1, 1.2, 0.5)$ and $\{\epsilon_{l,i}\}$ being i.i.d. $\mathcal{N}(0, \sigma^2/N)$. Here we compute each $\nabla \cdot [u_l e_k''(u_l) \nabla u_l]$ analytically since e_k and u_l are polynomials and trigonometric functions.

Regression from the self-test loss function. As studied in Section 4.1, the self-test loss function in (4.2) with Riemann sum approximation is

$$\begin{aligned} \mathcal{E}(\mathbf{c}) &= \frac{1}{NL} \sum_{l=1}^L \sum_{i=1}^N [h_{\mathbf{c}}''(u_l(x_i))^2 u_l(x_i) |\nabla u_l(x_i)|^2 - 2 f_l(x_i) h_{\mathbf{c}}'(u_l(x_i))] \\ &= \mathbf{c}^\top \mathbf{A} \mathbf{c} - 2 \mathbf{c}^\top \mathbf{b}, \end{aligned}$$

where the normal matrix $\mathbf{A} = (A_{k,m})$ and normal vector $\mathbf{b}$ defined by

$$\begin{aligned} A_{k,m} &= \frac{1}{NL} \sum_{l=1}^L \sum_{i=1}^N u_l(x_i) |\nabla u_l(x_i)|^2 e_k''(u_l(x_i)) e_m''(u_l(x_i)), \quad 1 \leq k, m \leq n_c \\ b_k &= \frac{1}{NL} \sum_{l=1}^L \sum_{i=1}^N f_l(x_i) e_k'(u_l(x_i)), \quad 1 \leq k \leq n_c. \end{aligned}$$

The estimator is then solved by least squares regression,

$$\hat{h}(s) = \sum_{k=1}^{n_c} \hat{c}_k e_k(s), \quad \hat{\mathbf{c}} = \mathbf{A}^{-1} \mathbf{b}.$$

We compare $\hat{h}(s)$ (denoted by “stLoss-estimator”) with an estimator using the strong-form equation (denoted by “Strong-estimator”). The strong form estimator has coefficient $\hat{\mathbf{c}}^s = (\mathbf{A}^s)^{-1} \mathbf{b}^s$, where the normal matrix $\mathbf{A}^s$ and normal vector $\mathbf{b}^s$ have entries

$$\begin{aligned} A_{k,m}^s &= \frac{1}{NL} \sum_{l=1}^L \sum_{i=1}^N \nabla \cdot [u_l e_k''(u_l) \nabla u_l](x_i) \nabla \cdot [u_l e_m''(u_l) \nabla u_l](x_i), \\ b_k^s &= \frac{1}{NL} \sum_{l=1}^L \sum_{i=1}^N f_l(x_i) \nabla \cdot [u_l e_k''(u_l) \nabla u_l](x_i). \end{aligned}$$

Thus, the strong form estimator uses the second-order derivatives of u , while the weak form estimator uses only the first-order derivatives.

In the computation of both estimators, the derivatives are approximated by finite difference using the Savitzky-Golay filter with polynomial degree 3 and window size 11 (see, e.g., [38]). The difference between the two is that the Strong-estimator requires an additional finite difference approximation for the divergence term, whereas the stLoss-estimator uses the Riemann sum to approximate the integration.

Figure 1 shows that the stLoss-estimator significantly outperforms the Strong-estimator in typical simulations (Left) and in the convergence of relative $L_{\rho_1}^2$ error ($\|h_*'' - \hat{h}''\|_{L_{\rho_1}^2}^2$)

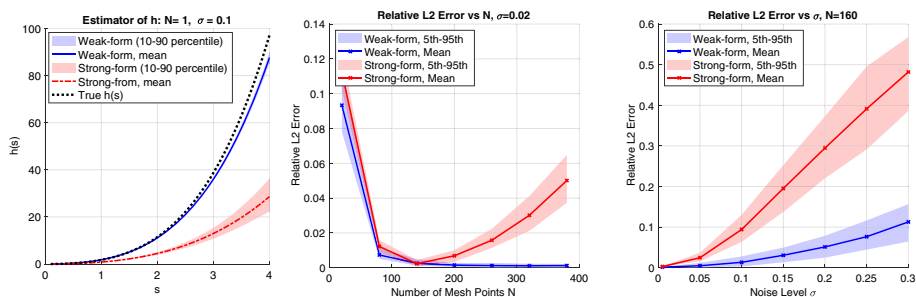


Fig. 1 Estimators from self-test loss function (“stLoss”) vs estimators from strong-form equation (“Strong”). **Left:** estimators in a typical set of 100 simulations with $N = 400$ and $\sigma = 0.1$. **Middle-Right:** Relative $L^2_{\rho_1}$ errors vs N and σ

as mesh size N increases or as the noise level σ vanishes (Middle-Right). Note that as N increases, the strong-estimator does not converge due to the noise being amplified by the additional finite difference approximation (recall that the noise decays at the rate $O(\frac{1}{\sqrt{N}})$) while the Δx in finite difference has an order of $O(\frac{1}{N})$.

Here, for each parameter set of (N, σ) , the percentiles are computed from 100 independent simulations with randomly generated noise; the empirical measure ρ_1 in (4.3) is computed from data by Riemann sum approximation of the integral and finite difference approximation of the derivatives. In the typical simulation (Left), the condition numbers of the normal matrix $\mathbf{A}$ are in the range [30, 40], indicating the well-posedness of the inverse problem.

In summary, the example shows that the estimator using our self-test loss function based on the weak-form equation can tolerate a rougher spatial mesh and larger-scale noise in the data than a strong-form-based estimator.

Application to weak SINDy. We apply our self-test loss function within the weak SINDy framework of [31] to estimate the sparse parametric diffusion rate h in (5.2) from data. Specifically, we compare our self-testing functions with random Gaussian test functions $\{\psi_m\}_{m=1}^M$ designed following the strategy in [31], namely, tailoring them to the noise level and spectral properties of the data. This design keeps the test set computationally feasible while avoiding the curse of dimensionality that affects more structured families. The Gaussian test functions have centers sampled uniformly in $[0, 1]^d$ (with $d = 2$) and bandwidths $\eta \in \{0.025, 0.1, 0.4\}$.

We assume the true coefficient for $h_c(s)$ is given by

$$\mathbf{c} = (c_1, c_2, c_3, c_4, c_5) = (1, 0, 2, 0, 0).$$

The data is generated on the discrete mesh, and for $d \geq 1$, we consider the data to be the tensor product $u_l(x) = \sin(\pi l x_1) \cdots \sin(\pi l x_d)$ evaluated over a discrete mesh with $N = 100$ grids in each dimension, where $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $1 \leq l \leq L$ with $L = 2$. The data u_l and f_l at each mesh point are polluted with additive Gaussian noise of variance $\sigma = 0.05$. Since the true parameter is known to be sparse, we use the modified sequential-thresholding least-squares (MSTLS) as introduced in [31] to promote sparsity in the estimation.

We report the estimation error as a function of M over 10 independent simulations in Figure 2. In the self-test setting, we always use $L = 2$ test functions, so the corresponding error is independent of M . As M increases, the error obtained with random Gaussian test

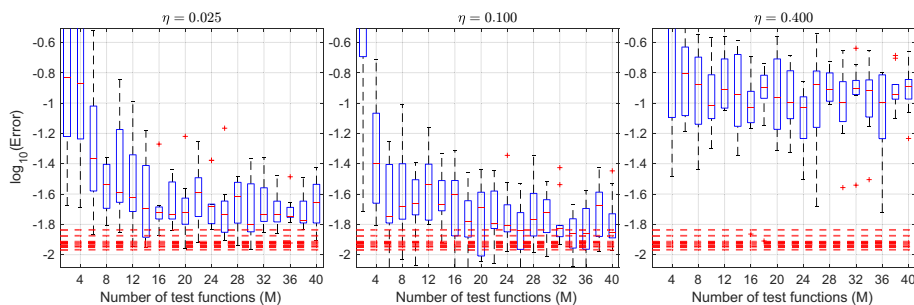


Fig. 2 Comparison between self-testing functions and random Gaussian test functions in the weak SINDy framework. The boxplots show the distribution of estimation errors over 10 independent simulations using M random Gaussian test functions, with bandwidths $\eta \in \{0.025, 0.1, 0.4\}$. The red dashed lines indicate the error obtained using $L = 2$ self-testing functions, which do not depend on M . As M increases, the error for Gaussian test functions decreases, and the best performance occurs when $\eta = 0.1$, close to the noise level $\sigma = 0.05$, at which point the errors approach those of the self-test formulation. Even in this near-optimal setting, self-testing functions yield consistently lower errors than the random Gaussian tests

functions decreases. The best performance occurs when the Gaussian bandwidth $\eta = 0.1$ is close to the noise level $\sigma = 0.05$, in which case the errors approach those achieved by the self-testing functions. Even in this near-optimal regime, the self-testing functions still outperform the random Gaussian test functions. This example highlights the advantage of the proposed self-test framework.

Separately, additional experiments (not shown) suggest that self-testing functions yield an estimation error on the order of the noise level. In contrast, with sufficiently large M , random Gaussian test functions can produce errors below the noise level. This indicates that, with an appropriate choice of bandwidth η , the random Gaussian test functions may effectively filter out the noise.

5.2 Non-Parametric Estimation of Interaction Kernel

Next, we consider estimating an interaction kernel, the derivative of a radial interaction potential, in the aggregation operator. We will compare strong-form and weak-form estimators with respect to their tolerance to observation noise.

Specifically, consider the estimation of the function $\phi : [0, 2] \rightarrow \mathbb{R}$, which is the derivative of the radial interaction potential Φ with $\nabla\Phi(x) = \phi(|x|)\frac{x}{|x|}$, in the aggregation operator

$$R_\phi[u] = -\nabla(u\nabla\Phi * u) = f,$$

from data consisting of noisy input-output function pairs at discrete meshes in 1D:

$$\{(u_k^o(x_j) := u_k(x_j) + \epsilon_{kj}^u, f_k^o(x_j) = f_k(x_j) + \epsilon_{kj}^f)\}_{j=1, k=1}^{n_x, n_k}, \quad (5.3)$$

where $\epsilon_{kj}^u, \epsilon_{kj}^f \sim \mathcal{N}(0, \sigma^2)$. Here, the spatial meshes $\{x_j\}$ are uniform on $\Omega = (0, 10)$ satisfying $x_j - x_{j-1} = \Delta x = 0.01$ for all j , and the noises $\{\epsilon_{kj}^u, \epsilon_{kj}^f\}_{k,j=1}^{n_k, n_x}$ are independent and identically distributed Gaussian $\mathcal{N}(0, \sigma^2)$ random variables with standard deviation σ . The functions $\{u_k\}$ are $u_k(x) = \sin(\pi(x - (2k + 1)))\mathbf{1}_{\{|x - (2k + 1)| < 1.5\}}$ for $1 \leq k \leq n_k = 3$. They are in $C_c^1(\Omega)$, so we can use integration by parts in the weak form and compute $f_k(x_j)$ using the strong form operator, i.e., we compute the analytical form of the integrand in the

following integral,

$$f_k(x) = - \int_0^2 \phi_*(|y|) \text{sign}(y) \partial_x [u_k(x-y)u_k(x)] dy,$$

where the integral is computed using the adaptive Gauss-Kronrod quadrature [39]. In our tests, we set $\phi_*(r) = r^2 \mathbf{1}_{[0,1]}(r)$. Figure 3(a) shows the data pairs.

The above equation is the mean-field equation (2.6) with $V = 0$ if $f = \partial_t u - v \Delta u$. In this case, the nonparametric estimation of ϕ has been studied in [23, 24]. Here, we focus on the aggregation operator without the diffusion term.

In the following, we derive the least squares regression of ϕ using the self-test loss function. We first write the self-test loss function in the continuum, then approximate it by the discrete data and write the least squares estimator of ϕ .

The self-test loss function in continuum. Using the self-testing function $v_\phi[u_k] = \Phi * u_k$ for each input-output pair (u_k, f_k) , and applying integration by parts, we obtain the self-test loss function

$$\mathcal{E}_{\mathcal{D}}(\phi) = \sum_{k=1}^{n_k} \int_{\mathbb{R}} u_k |\nabla \Phi * u_k|^2 dx - 2 \int_{\mathbb{R}} f_k(x) \Phi * u_k(x) dx.$$

Denote $F_{f,u}(r) := - \sum_{k=1}^{n_k} \int_0^{10} F_k(x) u_k(x) [u_k(x-r) - u_k(x+r)] dx$ with $F_k(x) := \int_0^x f_k(y) dy$. We can write the loss function as (see Appendix 7.3 for a derivation)

$$\mathcal{E}_{\mathcal{D}}(\phi) = \int \int \phi(r) \phi(s) \overline{G}(r, s) \dot{\rho}(r) \dot{\rho}(s) dr ds - 2 \int_0^2 \phi(r) F_{f,u}(r) dr, \quad (5.4)$$

where the density of the exploration measure $\dot{\rho}$ is defined as

$$\dot{\rho}(r) := \frac{1}{Z} \sum_{k=1}^{n_k} \int_{\mathbb{R}} \sqrt{u_k(x)} |\delta u_k(x, r)| dx \quad \text{with } \delta u_k(x, r) := u_k(x-r) - u_k(x+r) \quad (5.5)$$

with Z being a normalizing constant. Here, the integral kernel $\overline{G}$ is defined by

$$\overline{G}(r, s) := \frac{G(r, s)}{\dot{\rho}(r) \dot{\rho}(s)} \mathbf{1}_{\{\dot{\rho}(r) \dot{\rho}(s) > 0\}} \quad \text{with } G(r, s) := \sum_{k=1}^{n_r} \int_{\mathbb{R}} u_k(x) \delta u_k(x, r) \delta u_k(x, s) dx. \quad (5.6)$$

These integrals are well-defined since $\{u_k(x)\}$ are uniformly bounded with compact support.

Least squares regression from empirical loss function. Given the discrete data in (5.3) on the mesh $\{x_j\}$, we can obtain a uniform mesh $\{r_l = l \Delta x\}_{l=1}^{n_r}$ on $[0, 2]$ for the independent variable of ϕ . Representing ϕ by a linear combination of piecewise constant functions $\phi(r) = \sum_{l=1}^{n_r} c_l \mathbf{1}_{[r_l, r_{l+1}]}(r)$, our task is to estimate the coefficient vector $\mathbf{c} = (c_1, \dots, c_{n_r})^\top \in \mathbb{R}^{n_r \times 1}$. Approximating the loss function in (5.4) by Riemann sum using the noisy data and the above piecewise constant ϕ , we obtain an empirical loss function that is quadratic in $\mathbf{c}$:

$$\widehat{\mathcal{E}}_{\mathcal{D}}(\mathbf{c}) = \mathbf{c}^\top \mathbf{A} \mathbf{c} - 2 \mathbf{c}^\top \mathbf{b} + C,$$

where $\mathbf{A} \in \mathbb{R}^{n_r \times n_r}$ and $\mathbf{b} \in \mathbb{R}^{n_r \times 1}$ are the normal matrix and vectors and C is a constant term independent of $\mathbf{c}$. The entries of $\mathbf{A}$ and $\mathbf{b}$ are

$$\begin{aligned} \mathbf{A}(l, l') &= \mathbf{G}(l, l') \approx \int \int \mathbf{1}_{[r_l, r_{l+1}]}(r) \mathbf{1}_{[r_{l'}, r_{l'+1}]}(s) G(r, s) dr ds \\ \mathbf{b}(l) &= -\mathbf{g}^\top \widetilde{\mathbf{f}} \Delta r \Delta x \approx \int \mathbf{1}_{[r_l, r_{l+1}]}(r) F_{f,u}(r) dr, \end{aligned}$$

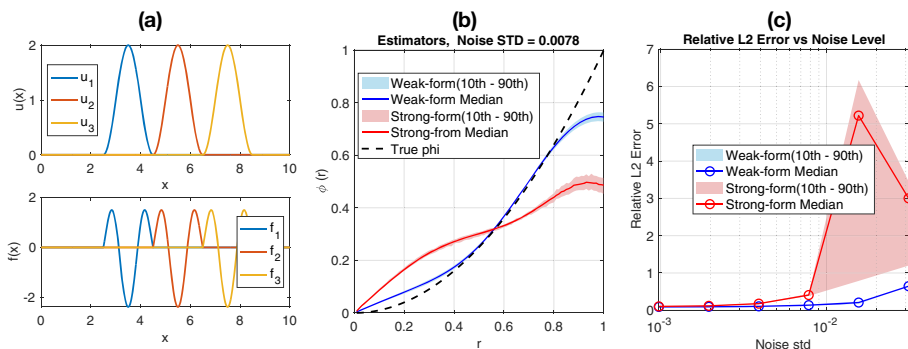


Fig. 3 Estimators using weak (“Weak-form”) vs strong-form (“Strong-form”) equation. **(a):** Dataset $\{(u_k, f_k)\}_{k=1}^3$. **(b):** Estimators (in percentiles) in a typical set of 100 simulations with noise level $\sigma = 0.0078$. **(c)** Relative L^2 error v.s. noise level σ . Weak-form estimators are more robust to large noise than those based on the strong-form

where we denote $\mathbf{G} = \mathbf{g}^\top \mathbf{g} \Delta x (\Delta r)^2 \in \mathbb{R}^{n_r \times n_r}$ with $\mathbf{g} = (\sqrt{|u_k^o(x_j)|} \delta u_k^o(x_j, r_l)) \in \mathbb{R}^{n_k n_x \times n_r}$ and $\tilde{\mathbf{f}} = (\sum_{i=1}^j f_k^o(x_i) u_k^o(x_j) \Delta x)_{k,j=1}^{n_k, n_x} \in \mathbb{R}^{n_k n_x \times 1}$.

The estimator is then solved with Tikhonov regularization:

$$\hat{\phi}(r) = \sum_{l=1}^{n_r} \hat{c}_l \mathbf{1}_{[r_l, r_{l+1}]}(r), \quad (\hat{c}_1, \dots, \hat{c}_{n_r})^\top := \hat{\mathbf{c}}_{\lambda_*} = (\mathbf{A} + \lambda_* \mathbf{I})^{-1} \mathbf{b}$$

with the hyperparameter $\lambda_* > 0$ selected by the L-curve method [16]. Due to the additive noise in u_k^o , the smallest eigenvalue of the normal matrix $\mathbf{A}$ is bounded below by a constant that scales with σ^2 . Thus, the noise prevents $\mathbf{A}$ from being severely ill-conditioned, and the regularization mainly acts as a filter of the noise. Here we regularize using norm $\|\mathbf{c}\|_{\mathbb{R}^{n_r}}^2 = \mathbf{c}^\top \mathbf{I} \mathbf{c}$, and we leave it in future work to investigate other norms, such as the L_ρ^2 -norm or the data-adaptive RKHS norm of ϕ in [26].

Also, we compute the exploration measure as $\boldsymbol{\rho} = (\dot{\rho}(r_1), \dots, \dot{\rho}(r_{n_r})) \in \mathbb{R}^{n_r \times 1}$ with $\dot{\rho}(r_l) = \frac{1}{Z} \sum_{k=1}^{n_k} \sum_{i=1}^{n_x} \sqrt{|u_k^o(x_i)|} |\delta u_k^o(x_i, r_l)| \Delta x$. The L_ρ^2 norm of ϕ is then given by $\|\phi\|_{L_\rho^2}^2 = \sum_{l=1}^{n_r} c_l^2 \dot{\rho}(r_l)$.

Numerical results. We compare the estimators using the weak-form and strong-form equations. The strong-form estimator uses the Savitzky-Golay filter to compute derivatives. We compute the estimators from data with noise levels $\sigma = \{2^{-j}, j = 5, \dots, 10\}$. We make 100 independent simulations for each noise level, each with randomly sampled noise.

Figure 3 (b)-(c) reports the estimators and relative errors using the median, the 10th and 90th percentiles. In particular, (b) shows that the weak-form estimator is more accurate than the strong-form estimator when the noise level is $\sigma = 2^{-7} \approx 0.0078$. (c) shows that when the noise level is small, the strong-form estimator is as accurate as the weak-form, indicating the effectiveness of the Savitzky-Golay filter. Still, when the noise level is high, the strong-form estimator has larger errors than the weak-form estimator, due to the need to approximate derivatives using finite differences.

In summary, the weak-form estimator outperforms the strong-form estimator in terms of robustness to high levels of noise.

5.3 Neural Network Regression for Joint Estimation

This section considers the joint estimation of the interaction potential Φ and the potential V of the deterministic interacting particle system in Example 3 from sequential ensembles of unlabeled data. We use the self-test loss function in (2.12) for the weak form PDE of the empirical measures, as derived in Section 2.4.

Numerical settings. In our test, we set $M = 10$, $N = 30$, $d = 2$, and $t_l = l\Delta t$ with $\Delta t = 0.01$ and $L = 20$. The particle system is solved using the fourth-order Runge-Kutta method. The true interaction and external force potentials are given by

$$\Phi^*(x) = \cos(2x_1^2) + \cos(x_2), \quad V^*(x) = \exp\left(-\frac{3}{10}(\sin(2x_1^2) + \arctan(x_2))\right). \quad (5.7)$$

In the data in (1.11), the initial conditions $(X_{t_l}^{i,(m)}, 1 \leq i \leq N) \in \mathbb{R}^{Nd}$ are randomly sampled, with half of the samples drawn from the uniform distribution over $[-2, 2]^{Nd}$ and the other half drawn from a Gaussian mixture, so that the data are spread over a region. Here $d = 2$ and the Gaussian mixture is the product measure of the distribution $0.6 \times \mathcal{N}(\mu_1, \Sigma_1) + 0.4 \times \mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^2$, where μ_1 are sampled from a uniform distribution on $[0, 2.5]^2$ and μ_2 are sampled from a uniform distribution on $[-2.5, 0]^2$. The covariance matrices are fixed to be $\Sigma_1 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.4 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. In this setting, the distribution of the particles is concentrated in the first and third quadrants, as shown in Figure 4f.

Regression via neural network approximation. We use neural networks to approximate both the interaction and external force potentials. To approximate the interaction and external force potentials, we use two four-layer fully connected neural networks with sigmoid and ReLU activation functions. In particular, we enforce symmetry by setting $\Phi(x) = \tilde{\Phi}(x) + \tilde{\Phi}(-x)$, where $\tilde{\Phi}$ is the neural network approximation. This constraint resolves the identifiability issue in Proposition 5 and in [47], where different pairs (Φ, V) and $(\Phi + \mathbf{c} \cdot x, V - \mathbf{c} \cdot x)$ produce the same value of the loss function, since $\Phi + \mathbf{c} \cdot x$ is only symmetric when $\mathbf{c} = 0$ if Φ is symmetric.

Optimization is performed using the Nesterov-accelerated Adaptive Moment Estimation NAdam method, which combines Adam's adaptive learning rates with Nesterov's lookahead mechanism to improve convergence and optimization efficiency [8], with a learning rate adjustment ReduceLROnPlateau, which reduces the learning rate when a monitored metric stops improving, helping to fine-tune optimization and avoid overfitting.

The training process is presented in Figure 5. The initial step size is set to be $\eta = 0.05$, and it is reduced to 0.1η whenever the loss stops reducing. The final minimized loss is -0.001309. Note that our self-test loss (2.1) is the quadratic (2.3) minus a constant, where the constant is related to the true functions. The true constant in this example is 0.001377, which suggests that the quadratic loss has been minimized to 6.69×10^{-5} .

Figure 4 presents the learned potentials. Figures 4a and 4d show the true and estimated interaction and force potentials, and the differences of their gradients are presented in Figures 4b and 4e. The estimators are accurate over the regions where data is concentrated, i.e., the large valued regions of the exploration measures, ρ_2 as in (4.6) for V and ρ_3 as in (4.11) for Φ , as shown in Figures 4c and 4f, respectively. These empirical measures are relatively rough since they are estimated from about $MNL = 4000$ and $MN^2L = 80000$ data samples for ρ_2 and ρ_3 , respectively. The final estimation error is $\|\nabla\Phi - \nabla\Phi^*\|_{L^2_{\rho_2}} = 0.5855$ and $\|\nabla V - \nabla V^*\|_{L^2_{\rho_2}} = 0.1746$.

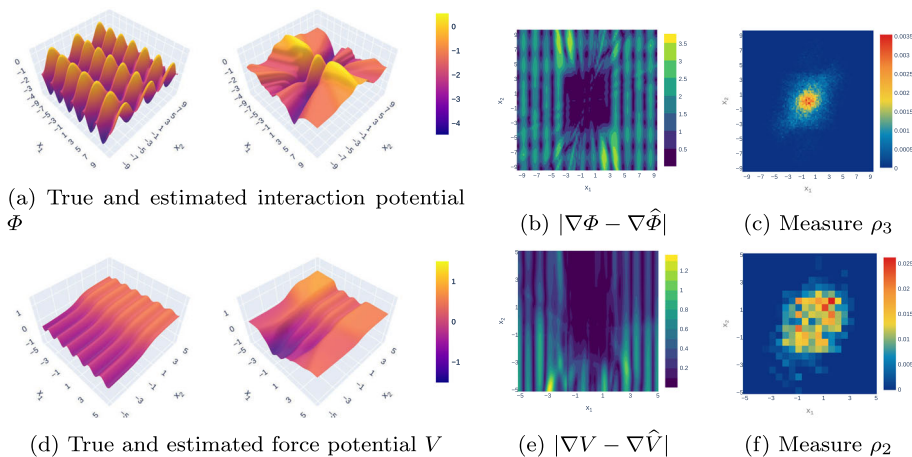


Fig. 4 Estimation result of the interaction and the force potential. The estimation results are accurate over the region where ρ_2 and ρ_3 are concentrated

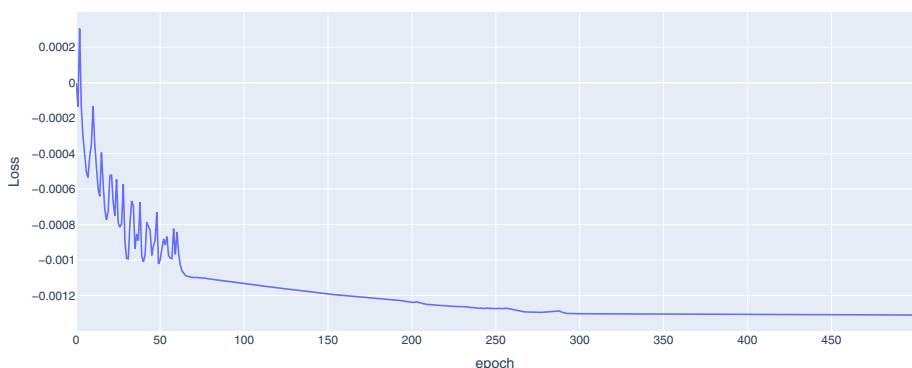


Fig. 5 Training process. We utilized the NAdam optimizer and adjusted the learning rate when the loss plateaued. The initial oscillation is attributed to the reduction in the learning rate. Toward the end, the loss decreases more gradually because the learning rate is significantly lower than at the beginning. The final optimized loss value is -0.001309, corresponding to a normalized quadratic loss of 6.69×10^{-5}

To summarize, we overcome the challenge of unlabeled ensemble data without trajectory information by constructing a self-test loss function based on the weak-form equation of the empirical distributions. This self-test loss function is suitable for ensemble unlabeled data and neural network regression.

6 Conclusion

Discrete, noisy data pose substantial challenges for learning differential operators in PDEs and gradient flow systems. A standard approach is to construct loss functions based on weak-form equations, which avoids the large errors inherent in approximating high-order derivatives. However, this introduces the challenge of selecting suitable test functions.

This paper introduced a novel framework for constructing loss functions, called self-test loss functions. This method is designed for weak-form operators in PDEs and gradient flow systems. It applies to operators that depend linearly on the (function-valued) parameter to be estimated. By leveraging parameter—and data-dependent test functions, our approach automates the construction of loss functions and addresses the issue of test function selection.

The self-test loss function exhibits appealing theoretical and computational properties. It conserves energy in gradient flows and aligns with the expected log-likelihood ratio in stochastic differential equations. Furthermore, its quadratic structure enables a comprehensive analysis of the identifiability and well-posedness of the inverse problem. We demonstrate this by estimating the diffusion rate function, interaction potential, and kinetic potential in the aggregation-diffusion equation. Importantly, the self-test loss function supports the development of efficient parametric and nonparametric regression algorithms. Numerical experiments demonstrate that its minimizer is robust to noisy and discrete data, highlighting its practical utility and potential for broader applications.

7 Proofs and Derivations

7.1 Proofs for Section 3

Proof of Theorem 3.2 Part (a). Since ϕ_* is the true parameter, it satisfies the weak form of gradient flow $\partial_t u = -A_u \frac{\delta E_{\phi_*}}{\delta u}$. Applying a test function $\frac{\delta E_{\phi}}{\delta u}$ for any ϕ such that $E_{\phi}[u] < \infty$, we obtain

$$\frac{dE_{\phi}[u]}{dt} = \left\langle \partial_t u, \frac{\delta E_{\phi}}{\delta u} \right\rangle = - \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi}}{\delta u} \right\rangle, \quad \forall t \in [0, T].$$

Integrating in time, we obtain

$$E_{\phi}(u(T, \cdot)) - E_{\phi}(u(0, \cdot)) = \int_0^T \frac{dE_{\phi}[u]}{dt} dt = - \int_0^T \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi}}{\delta u} \right\rangle dt.$$

Then, using the linearity of $\frac{\delta E_{\phi}}{\delta u}$ in ϕ due to Assumption 3.1, we write $\mathcal{E}_{u[0,T]}(\phi)$ in (3.4) as

$$\begin{aligned} \mathcal{E}_{u[0,T]}(\phi) &= -2 \int_0^T \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi}}{\delta u} \right\rangle dt + \int_0^T \left\langle A_u \frac{\delta E_{\phi}}{\delta u}, \frac{\delta E_{\phi}}{\delta u} \right\rangle dt \\ &= \int_0^T \left\langle A_u \frac{\delta E_{\phi - \phi_*}}{\delta u}, \frac{\delta E_{\phi - \phi_*}}{\delta u} \right\rangle dt - \int_0^T \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi_*}}{\delta u} \right\rangle dt. \end{aligned} \quad (7.1)$$

From $\langle A_u \xi, \xi \rangle \geq 0$ in (3.3), the first term is non-negative. Thus, we know that

$$\mathcal{E}_{u[0,T]}(\phi) \geq \mathcal{E}_{u[0,T]}(\phi_*) = - \int_0^T \left\langle A_u \frac{\delta E_{\phi_*}}{\delta u}, \frac{\delta E_{\phi_*}}{\delta u} \right\rangle dt$$

and ϕ_* is a minimizer of $\mathcal{E}_{u[0,T]}(\phi)$.

Part (b). It follows from (7.1) that ϕ_* is the unique minimizer in $\mathcal{H}$ if (3.5) holds.

Part (c). Since ϕ_0 is a minimizer of $\mathcal{E}_{u[0,T]}(\phi)$ and E_{ϕ} is linear in ϕ , we have, for any ψ ,

$$\begin{aligned} 0 &= \frac{d}{d\epsilon} \mathcal{E}_{u[0,T]}(\phi_0 + \epsilon \psi) = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{E}_{u[0,T]}(\phi_0 + \epsilon \psi) - \mathcal{E}_{u[0,T]}(\phi_0)}{\epsilon} \\ &= 2[E_{\psi}(u(T, \cdot)) - E_{\psi}(u(0, \cdot))] + 2 \int_0^T \left\langle A_u \frac{\delta E_{\phi_0}}{\delta u}, \frac{\delta E_{\psi}}{\delta u} \right\rangle dt. \end{aligned}$$

Taking $\psi = \phi_0$, we obtain (3.6). $\square$

Proof of Theorem 3.3 The Fokker-Planck equation of the McKean-Vlasov SDE is (2.6). The self-test loss function for estimating (V, Φ) using its weak form is given in (2.9), which reads

$$\mathcal{E}_{u_{[0,T]}}(\Phi, V) := \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} [u |\nabla \Phi * u + \nabla V|^2 - 2(\partial_t u - v \Delta u)(\Phi * u + V)] dx dt.$$

On the other hand, by Girsanov Theorem (see e.g., [34]), the negative log-likelihood ratio for $\bar{X}_{[0,T]}$ is

$$\mathcal{E}_{\bar{X}_{[0,T]}}(\phi) = -\ln \frac{d\mathbb{P}_\phi}{d\mathbb{P}_0} = \frac{1}{2v} \int_0^T \left(|\nabla \Phi * u + \nabla V|(\bar{X}_t)|^2 dt - 2[\nabla \Phi * u + \nabla V](\bar{X}_t), d\bar{X}_t \right),$$

where $\mathbb{P}_\phi$ and $\mathbb{P}_0$ are the distributions of the path under the SDE with parameters $\phi = (V, \Phi)$ and $V = \Phi = 0$, respectively. Taking expectation and using the fact that $\bar{X}_t \sim u(\cdot, t)$,

$$\mathbb{E}[\mathcal{E}_{\bar{X}_{[0,T]}}(\phi)] = \frac{1}{2v} \int_0^T \int_{\mathbb{R}^d} [|\nabla \Phi * u + \nabla V|^2 u dx - 2\mathbb{E}[(\nabla \Phi * u + \nabla V)(\bar{X}_t), d\bar{X}_t]] dt.$$

To compute the above expectation, using $d\bar{X}_t$ from the SDE with the fact that the martingale term has expectation 0 and applying integration by parts, we have

$$\begin{aligned} \mathbb{E}[(\nabla \Phi * u + \nabla V)(\bar{X}_t), d\bar{X}_t] &= \mathbb{E}[(\nabla(\Phi * u + V)(\bar{X}_t), -\nabla[V_* + \Phi_* * u](\bar{X}_t))] \\ &= \int_{\mathbb{R}^d} \langle \nabla(\Phi * u + V), -u \nabla[V_* + \Phi_* * u] \rangle dx \\ &= \int_{\mathbb{R}^d} (\Phi * u + V) \nabla \cdot [u \nabla(V_* + \Phi_* * u)] dx = \int_{\mathbb{R}^d} (\Phi * u + V)(\partial_t u - v \Delta u) dx, \end{aligned}$$

where the last equation follows from the Fokker-Planck equation (2.6) with parameters (V_*, Φ_*) . Combining the above two equations, we have $\mathcal{E}_{u_{[0,T]}}(\Phi, V) = \mathcal{E}_{\bar{X}_{[0,T]}}(\phi)$. $\square$

7.2 Proofs for Section 4

Proof of Proposition 2 Recall that in (4.1), $\tilde{f} = -\nabla \cdot [u \nabla [v h'_*(u) + (\Phi_* - \Phi) * u + V_* - V]] =: -\nabla \cdot [u \nabla F[u]]$, where we set $F[u] = v h'_*(u) + (\Phi_* - \Phi) * u + V_* - V$. Then,

$$\begin{aligned} \langle \tilde{f}, v_h[u] \rangle &= \langle -\nabla \cdot [u \nabla F[u]], h'(u) \rangle = \int_{\mathbb{R}^d} u(x) \nabla F[u](x) \cdot \nabla u(x) h''(u(x)) dx \\ &\leq \left(\int_{\mathbb{R}^d} u(x) |\nabla F[u](x)|^2 dx \right)^{1/2} \left(\int_{\mathbb{R}^d} u(x) |\nabla u(x)|^2 h''(u(x))^2 dx \right)^{1/2} < +\infty. \end{aligned}$$

Thus, the Riesz representation theorem gives a data-dependent $h_{\mathcal{D}} \in L^2_{\rho_1}$ with ρ_1 defined in (4.3) such that

$$\begin{aligned} \sum_{l=1}^L \langle \tilde{f}_l, v_h[u_l] \rangle &= \sum_{l=1}^L \langle -\nabla \cdot [u_l \nabla F[u_l]], h'(u_l) \rangle \\ &= \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \nabla F[u_l] \cdot \nabla u_l h''(u_l(x)) dx =: \langle h_{\mathcal{D}}, h'' \rangle_{L^2_{\rho_1}}. \end{aligned}$$

Then, we can write the self-test loss function as

$$\begin{aligned}\mathcal{E}_1(h'') &= \sum_{l=1}^L \langle R_h[u_l] - 2\tilde{f}_l, v_h[u_l] \rangle + C_0 \\ &= \|h''\|_{L^2_{\rho_1}}^2 - 2\langle h'', h_D \rangle_{L^2_{\rho_1}} + C_0.\end{aligned}$$

The Fréchet derivative of $\mathcal{E}_1$ in terms of the variable h'' is $D_{h''}\mathcal{E}_1(h'') = 2h'' - 2h_D$. Thus, the minimizer of $\mathcal{E}_1$ is unique and

$$\hat{h}'' = \arg \min_{h'' \in L^2_{\rho_1}} \mathcal{E}_1(h'') = I^{-1}h_D$$

with I being the identity operator on $L^2_{\rho_1}$. Thus, this inverse problem is well-posed. $\square$

Proof of Proposition 3 First, recall that in (4.4), $\tilde{f} = -\nabla \cdot [u \nabla [v h'_*(u) - v h'(u) + (\Phi_* - \Phi) * u + V_*]] := -\nabla \cdot [u \nabla F[u]]$, where we set $F[u] = v h'_*(u) - v h'_*(u) + (\Phi_* - \Phi) * u + V_*$. Thus, the linear term in the loss function is

$$\sum_{l=1}^L \langle \tilde{f}_l, v_V[u_l] \rangle = \sum_{l=1}^L \langle -\nabla \cdot [u_l \nabla F[u_l]], V \rangle = \sum_{l=1}^L \langle u_l \nabla F[u_l] \cdot \nabla V \rangle =: \langle \vec{V}_D, \nabla V \rangle_{L^2_{\rho_2}},$$

where $\vec{V}_D \in L^2_{\rho_2}(\mathbb{R}^d; \mathbb{R}^d)$ with ρ_2 defined in (4.6) by the Riesz representation theorem. In particular, we have $\vec{V}_D = \nabla F[u_l]$ when $F[u_l]$ is independent of l (e.g., when $L = 1$). Then, we can write the self-test loss function as

$$\begin{aligned}\mathcal{E}_2(\nabla V) &= \sum_{l=1}^L \langle R_V[u_l] - 2\tilde{f}_l, V \rangle \\ &= \|\nabla V\|_{L^2_{\rho_2}}^2 - 2\langle \nabla V, \vec{V}_D \rangle_{L^2_{\rho_2}} + C_0.\end{aligned}$$

Regarding $\mathcal{E}_2(\nabla V)$ as a functional of $\mathbf{v} = \nabla V \in L^2_{\rho_2}(\mathbb{R}^d; \mathbb{R}^d)$, we define $\mathcal{E}_2(\mathbf{v}) = \|\mathbf{v}\|_{L^2_{\rho_2}}^2 - 2\langle \mathbf{v}, \vec{V}_D \rangle_{L^2_{\rho_2}} + C_0$. The Fréchet derivative of $\mathcal{E}_2$ over $L^2_{\rho_2}(\mathbb{R}^d; \mathbb{R}^d)$ is $D_{\mathbf{v}}\mathcal{E}_2(\mathbf{v}) = 2(\mathbf{v} - \vec{V}_D)$.

Thus, the minimizer of $\mathcal{E}_2$, denoted as $\widehat{\nabla V}$, is unique and satisfies

$$\widehat{\nabla V} = \arg \min_{\nabla V \in L^2_{\rho_2}(\mathbb{R}^d; \mathbb{R}^d)} \mathcal{E}_2(\nabla V) = I^{-1}\vec{V}_D,$$

where I is the identity operator in $L^2_{\rho_2}(\mathbb{R}^d; \mathbb{R}^d)$, and the inverse problem is well-posed.

To identify V , we regard the self-test loss function as a functional from $\mathcal{H}_0$ to $\mathbb{R}$:

$$\tilde{\mathcal{E}}_2(V) := \mathcal{E}_2(\nabla V) = \|\nabla V\|_{L^2_{\rho_2}}^2 - 2 \left\langle \sum_{l=1}^L \tilde{f}_l, V \right\rangle_{L^2} + C_0.$$

Using Poincaré inequality (4.7), we have $\int_{\mathbb{R}^d} |V|^2 \rho_2 \, dx \leq c \int_{\mathbb{R}^d} |\nabla V|^2 \rho_2 \, dx$, where $c > 0$ is the Poincaré constant. This implies $\|V\|_{\mathcal{H}_0}^2 := \|V\|_{H^1_{\rho_2}}^2 \leq (1+c)\|\nabla V\|_{L^2_{\rho_2}}^2$. Combining this with Hölder's inequality for

$$\left| \left\langle \sum_{l=1}^L \tilde{f}_l, V \right\rangle_{L^2_{\rho_2}} \right| = |\langle \nabla V, \vec{V}_D \rangle_{L^2_{\rho_2}}| \leq \frac{1}{4(1+c)} \|\nabla V\|_{L^2_{\rho_2}}^2 + 4(1+c) \|\vec{V}_D\|_{L^2_{\rho_2}}^2,$$

so we have

$$\tilde{\mathcal{E}}_2(V) \geq \frac{1}{2(1+c)} \|V\|_{H_{\rho_2}^1}^2 + C_0 - 8(1+c) \|\vec{V}_{\mathcal{D}}\|_{L_{\rho_2}^2}^2.$$

Hence, the functional $\tilde{\mathcal{E}}_2(V)$ is uniformly convex on $\mathcal{H}_0$, and it has a unique minimizer in $\mathcal{H}_0$.

If $\widehat{V}$ minimizes $\tilde{\mathcal{E}}_2(V)$, the first variation (Gateaux derivative) of $\tilde{\mathcal{E}}_2(V)$ is

$$\left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \int |\nabla(\widehat{V} + \varepsilon \tilde{V})|^2 \rho_2 - 2(\widehat{V} + \varepsilon \tilde{V}) \sum_{l=1}^L \tilde{f}_l \, dx = 2 \int \left(\nabla \widehat{V} \nabla \tilde{V} \rho_2 - \sum_{l=1}^L \tilde{f}_l \tilde{V} \right) dx = 0$$

for any $\tilde{V} \in \mathcal{H}_0$. Hence, the minimizer $\widehat{V}$ satisfies (4.8). $\square$

Proof of Proposition 4 First, write the quadratic term in the loss function (4.10) as

$$\begin{aligned} & \sum_{l=1}^L \int_{\mathbb{R}^d} u_l(x) |\nabla \Phi * u_l(x)|^2 dx \\ &= \sum_{l=1}^L \int u_l(x) \int \nabla \Phi(y) u_l(x-y) dy \cdot \int \nabla \Phi(y') u_l(x-y') dy' dx \\ &= \int \int \langle \nabla \Phi(y), \nabla \Phi(y') \rangle_{\mathbb{R}^d} \left[\sum_{l=1}^L \int u_l(x) u_l(x-y) u_l(x-y') dx \right] dy dy' = \langle \nabla \Phi, L_{\overline{G}} \nabla \Phi \rangle_{L_{\rho_3}^2}, \end{aligned}$$

with $L_{\overline{G}}$ defined in (4.12) and ρ_3 defined in (4.11).

Second, the Riesz representation theorem gives a vector-valued function $\vec{\Phi}_{\mathcal{D}} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that the linear term in the loss function can be written as

$$\sum_{l=1}^L \int_{\mathbb{R}^d} u_l(x) \nabla F[u_l](x) \cdot \nabla \Phi * u_l(x) dx = \langle \vec{\Phi}_{\mathcal{D}}, \nabla \Phi \rangle_{L_{\rho_3}^2}.$$

Then, we can write the loss function in (4.10) as

$$\mathcal{E}_3(\nabla \Phi) = \langle \nabla \Phi, L_{\overline{G}} \nabla \Phi \rangle_{L_{\rho_3}^2} - 2 \langle \vec{\Phi}_{\mathcal{D}}, \nabla \Phi \rangle_{L_{\rho_3}^2} + C_0. \quad (7.2)$$

Regarding $\mathcal{E}_3$ as a functional in terms of $\nabla \Phi$, the Fréchet derivative of $\mathcal{E}_3$ is $D_{\nabla \Phi} \mathcal{E}_3(\nabla \Phi) = 2L_{\overline{G}} \nabla \Phi - 2\vec{\Phi}_{\mathcal{D}}$. Thus, the minimizer of $\mathcal{E}_3$ is unique in $\text{Null}(L_{\overline{G}})^\perp$ and

$$\widehat{\nabla \Phi} = \arg \min_{\nabla \Phi \in \text{Null}(L_{\overline{G}})^\perp \subset L_{\rho_3}^2} \mathcal{E}_3(\nabla \Phi) = L_{\overline{G}}^{-1} \vec{\Phi}_{\mathcal{D}}, \quad (7.3)$$

where $L_{\overline{G}}^{-1}$ is the pseudo-inverse of the operator $L_{\overline{G}}$. Since the operator $L_{\overline{G}}$ is compact, so $\text{Null}(L_{\overline{G}}) \neq \{0\}$ and the above inverse problem is ill-posed. $\square$

Proof of Proposition 5 We solve for the estimators by setting the Fréchet derivatives of the loss function to zero. We write the loss function in (4.13) as

$$\begin{aligned}\mathcal{E}(h'', \nabla V, \nabla \Phi) &= \|h''\|_{L_{\rho_1}^2}^2 + \|\nabla V\|_{L_{\rho_2}^2}^2 + \langle \nabla \Phi, L_{\bar{G}} \nabla \Phi \rangle_{L_{\rho_3}^2} \\ &\quad + 2 \sum_{l=1}^L \int_{\mathbb{R}^d} u_l v h''(u_l) \nabla u \cdot (\nabla V + \nabla \Phi * u) dx \\ &\quad + 2 \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \nabla V \cdot \nabla \Phi * u_l dx \\ &\quad - 2 \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \nabla v_{\phi_*}[u_l] \cdot \nabla [v h'(u_l) + \Phi * u_l + V] dx.\end{aligned}$$

Recall ρ_1, ρ_2, ρ_3 defined in (4.3), (4.6) and (4.11), respectively. We have that

$$\begin{aligned}\langle D_{h''} \mathcal{E}(h'', \nabla V, \nabla \Phi), g_1 \rangle_{L_{\rho_1}^2} &= \langle 2(h'' + M_{hV} \nabla V + M_{h\Phi} \nabla \Phi - h_{\mathcal{D}}), g_1 \rangle_{L_{\rho_1}^2}, \\ \langle D_{\nabla V} \mathcal{E}(h'', \nabla V, \nabla \Phi), \vec{g}_2 \rangle_{L_{\rho_2}^2} &= \langle 2(M_{Vh} h'' + \nabla V + M_{V\Phi} \nabla \Phi - \vec{V}_{\mathcal{D}}), \vec{g}_2 \rangle_{L_{\rho_2}^2}, \\ \langle D_{\nabla \Phi} \mathcal{E}(h'', \nabla V, \nabla \Phi), \vec{g}_3 \rangle_{L_{\rho_3}^2} &= \langle 2(M_{\Phi h} h'' + M_{\Phi V} \nabla V + L_{\bar{G}} \nabla \Phi - \vec{\Phi}_{\mathcal{D}}), \vec{g}_3 \rangle_{L_{\rho_3}^2},\end{aligned}$$

$\forall g_1 \in L_{\rho_1}^2, \vec{g}_2 \in L_{\rho_2}^2$, and $\vec{g}_3 \in L_{\rho_3}^2$. Here, the operators M_{ab} are defined from the cross-product terms in the loss function. For example,

$$\begin{aligned}\langle M_{hV} \nabla V, g_1 \rangle_{L_{\rho_1}^2} &= \sum_{l=1}^L \int_{\mathbb{R}^d} u_l v g_1 \nabla u \cdot \nabla V dx; \\ \langle M_{Vh} h'', \vec{g}_2 \rangle_{L_{\rho_2}^2} &= \sum_{l=1}^L \int_{\mathbb{R}^d} u_l v h'' \nabla u \cdot \vec{g}_2; \\ \langle M_{V\Phi} \nabla \Phi, \vec{g}_2 \rangle_{L_{\rho_2}^2} &= \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \vec{g}_2 \cdot \nabla \Phi * u_l dx; \\ \langle M_{\Phi V} \nabla V, \vec{g}_3 \rangle_{L_{\rho_3}^2} &= \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \nabla V \cdot \vec{g}_3 * u_l dx.\end{aligned}$$

In particular, since

$$\begin{aligned}\langle M_{hV} \vec{b}, g_1 \rangle_{L_{\rho_1}^2} &= \langle M_{Vh} g_1, \vec{b} \rangle_{L_{\rho_2}^2}, \quad \forall \vec{b} \in L_{\rho_2}^2, g_1 \in L_{\rho_1}^2, \\ \langle M_{\Phi V} \vec{g}_2, \vec{g}_3 \rangle_{L_{\rho_3}^2} &= \langle M_{V\Phi} \vec{g}_3, \vec{g}_2 \rangle_{L_{\rho_2}^2}, \quad \forall \vec{g}_2 \in L_{\rho_2}^2, \vec{g}_3 \in L_{\rho_3}^2,\end{aligned}$$

we have joint operators $M_{hV}^* = M_{Vh}$ and $M_{\Phi V}^* = M_{V\Phi}$ with operator norms satisfying $\|M_{hV}\| \leq 1$ and $\|M_{\Phi V}\| \leq \|L_{\bar{G}}\|^{1/2}$. Then, the joint estimator solves the system

$$\begin{pmatrix} I_{L_{\rho_1}^2} & M_{hV} & M_{h\Phi} \\ M_{Vh} & I_{L_{\rho_2}^2} & M_{V\Phi} \\ M_{\Phi h} & M_{\Phi V} & L_{\bar{G}} \end{pmatrix} \begin{pmatrix} h'' \\ \nabla V \\ \nabla \Phi \end{pmatrix} = \begin{pmatrix} h_{\mathcal{D}} \\ \vec{V}_{\mathcal{D}} \\ \vec{\Phi}_{\mathcal{D}} \end{pmatrix}. \quad (7.4)$$

The Hessian (the second variation) of the loss function is the operator on the left-hand-side of (7.4), and denote it by $A : L^2_{\rho_1}(\mathbb{R}^+) \otimes L^2_{\rho_2}(\mathbb{R}^d) \otimes L^2_{\rho_3}(\mathbb{R}^d) \rightarrow L^2_{\rho_1}(\mathbb{R}^+) \otimes L^2_{\rho_2}(\mathbb{R}^d) \otimes L^2_{\rho_3}(\mathbb{R}^d)$. The operator A is self-adjoint and semi-positive definite.

We show first that $\phi = (0, \mathbf{c}, -\mathbf{c})$ with a nonzero $\mathbf{c} \in \mathbb{R}^d$ is an eigenfunction of A corresponding to the zero eigenvalue. Note that by definition, $\langle M_V \phi \mathbf{c}, \mathbf{c} \rangle_{L^2_{\rho_2}} = \sum_{l=1}^L \int_{\mathbb{R}^d} u_l \mathbf{c} \cdot \mathbf{c} * u_l dx = \|\mathbf{c}\|^2$ and similarly, $\langle M_{\Phi V} \mathbf{c}, \mathbf{c} \rangle_{L^2_{\rho_3}} = \|\mathbf{c}\|^2$. Meanwhile, we have $L_{\overline{G}} \mathbf{c} = \int \int \frac{G(y, y')}{\rho_3(y)} \mathbf{c} dy' dy = \mathbf{c}$ by the definition of G . It follows that

$$\begin{aligned} \langle A\phi, \phi \rangle_{L^2_{\rho_1} \otimes L^2_{\rho_2} \otimes L^2_{\rho_3}} &= \left\langle \begin{pmatrix} M_h V \mathbf{c} - M_h \Phi \mathbf{c} \\ I_{L^2_{\rho_2}} \mathbf{c} - M_V \phi \mathbf{c} \\ M_{\Phi V} \mathbf{c} - L_{\overline{G}} \mathbf{c} \end{pmatrix}, \begin{pmatrix} 0 \\ \mathbf{c} \\ -\mathbf{c} \end{pmatrix} \right\rangle_{L^2_{\rho_1} \otimes L^2_{\rho_2} \otimes L^2_{\rho_3}} \\ &= \langle I_{L^2_{\rho_2}} \mathbf{c} - M_V \phi \mathbf{c}, \mathbf{c} \rangle_{L^2_{\rho_2}} - \langle M_{\Phi V} \mathbf{c} - L_{\overline{G}} \mathbf{c}, \mathbf{c} \rangle_{L^2_{\rho_3}} = 0. \end{aligned}$$

Lastly, note that for $\phi_n = (0, 0, \psi_n)$, where ψ_n is an eigenfunction of $L_{\overline{G}}$ such that $L_{\overline{G}} \psi_n = \lambda_n \psi_n$, we have $\langle A\phi_n, \phi_n \rangle = \lambda_n$, where $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$ since $L_{\overline{G}}$ is compact. Thus, the loss function is not uniformly convex, and the joint estimation is ill-posed. $\square$

7.3 Derivation Details for Section 5.2

Derivation of Eq.(5.4) Using the facts that $\nabla \Phi(|x|) = \phi(|x|) \frac{x}{|x|}$ and

$$\nabla \Phi * u(x) = \int_{\mathbb{R}} \phi(|y|) \frac{y}{|y|} u(x-y) dy = \int_0^\infty \phi(r) [u(x-r) - u(x+r)] dr,$$

along with the notation $\delta u(x, r; t)$ in (5.5), we can write the integrals as

$$\begin{aligned} \int_{\mathbb{R}} |u| |\nabla \Phi * u|^2 dx dt &= \int_0^\infty \int_0^\infty \phi(r) \phi(s) \int_{\mathbb{R}} u(x) \delta u(x, r) \delta u(x, s) dx dr ds dt, \\ &= \int_0^\infty \int_0^\infty \phi(r) \phi(s) G(r, s) dr ds \\ &= \int_0^\infty \int_0^\infty \phi(r) \phi(s) \overline{G}(r, s) \dot{\rho}(r) \dot{\rho}(s) dr ds. \end{aligned}$$

where the integral kernels $G, \overline{G} : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ are defined in (5.6).

Denote $F(x) = \int_0^x f(y) dy$. Integration by parts with $\Phi * u(10) = \Phi * u(0) = 0$ implies that

$$\begin{aligned} \int_{\mathbb{R}} f(x) \Phi * u(x) dx &= F(x) \Phi * u(x) \Big|_0^{10} - \int_0^2 \phi(r) \int_0^{10} F(x) [u(x-r) - u(x+r)] dx dr \\ &= \int_0^2 \phi(r) F_{f,u}(r) dr, \end{aligned}$$

where $F_{f,u}(r) := -\int_0^{10} F(x) [u(x-r) - u(x+r)] dx$. Combining the above equations, we obtain (5.4). $\square$

Data Availability The data used in this study are synthetic and were generated through simulations. Detailed descriptions of the simulation methodologies and parameters are provided in the manuscript to ensure reproducibility. All data sets and codes will be made available upon request.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

References

1. Bakry, D., Gentil, I., Ledoux, M.: Analysis and Geometry of Markov Diffusion Operators. Grundlehren der Mathematischen Wissenschaften, vol. 348. Springer, Cham (2014)
2. Bao, G., Ye, X., Zang, Y., Zhou, H.: Numerical solution of inverse problems by weak adversarial networks. *Inverse Prob.* **36**(11), 115003 (2020)
3. Carrillo, J., Craig, K., Yao, Y.: Aggregation-diffusion equations: dynamics, asymptotics, and singular limits. In *Active Particles, Volume 2*, pages 65–108. Springer, (2019)
4. Jose, A., Carrillo, G., Estrada-Rodriguez, L., Mikolas, S., Tang: Sparse identification of nonlocal interaction kernels in nonlinear gradient flow equations via partial inversion, (2024). arXiv preprint [arXiv:2402.06355](https://arxiv.org/abs/2402.06355)
5. Chada, N.K., Lang, Q., Lu, F., Wang, X.: A data-adaptive RKHS prior for Bayesian learning of kernels in operators. *J. Mach. Learn. Res.* **25**(317), 1–37 (2024)
6. Champion, K., Bethany Lusch, J., Kutz, N., Brunton, S.L.: Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci.* **116**(45), 22445–22451 (2019)
7. De Ryck, T., Mishra, S., Molinaro, R.: wpinns: Weak physics informed neural networks for approximating entropy solutions of hyperbolic conservation laws. *SIAM J. Numer. Anal.* **62**(2), 811–841 (2024)
8. Dozat, T.: Incorporating Nesterov momentum into Adam, (2016). (ICLR 2016 workshop)
9. E, W., Yu, B. The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems. *Commun. Math. Stat.* **6**(1), 1–12 (2018)
10. Gao, H., Zahr, M.J., Wang, J.: X Physics-informed graph neural galerkin networks: A unified framework for solving pde-governed forward and inverse problems. *Comput. Methods Appl. Mech. Eng.* **390**, 114502 (2022)
11. Gao, Y., Liu, J.-G., Lu, J., Marzuola, J.L.: Analysis of a continuum theory for broken bond crystal surface models with evaporation and deposition effects. *Nonlinearity* **33**(8), 3816 (2020)
12. Gao, Y., Liu, J.-G., Lu, X.Y.: Gradient flow approach to an exponential thin film equation: global existence and latent singularity. *ESAIM: Control, Optimisation and Calculus of Variations* **25**, 49 (2019)
13. Ghattas, O., Willcox, K.: Learning physics-based models from data: perspectives from inverse problems and model reduction. *Acta Numer* **30**, 445–554 (2021)
14. Gilpin, W., Huang, Y., Forger, D.B.: Learning dynamics from large biological data sets: machine learning meets systems biology. *Current Opinion in Systems Biology* **22**, 1–7 (2020)
15. Hamzi, B., Owhadi, H.: Learning dynamical systems from data: A simple cross-validation perspective, part i: Parametric kernel flows. *Physica D* **421**, 132817 (2021)
16. Hansen, P.C.: The L-curve and its use in the numerical treatment of inverse problems. In *Computational Inverse Problems in Electrocardiology*, ed. P. Johnston, *Advances in Computational Bioengineering*, pages 119–142. WIT Press, (2000)
17. Hu, Z., Liu, C., Wang, Y., Xu, Z. Energetic variational neural network discretizations of gradient flows. *SIAM J. Sci. Comput.* **46**(4), A2528–A2556 (2024)
18. Isakov, V.: *Inverse problems for partial differential equations*, volume 127. Springer, (2006)
19. Jabin, P.-E., Wang, Z.: Mean field limit and propagation of chaos for Vlasov systems with bounded forces. *J. Funct. Anal.* **271**(12), 3588–3627 (2016)
20. Jabin, P.E., Wang, Z.: Mean field limit for stochastic particle systems. In *Active Particles, Volume 1*, pages 379–402. Springer, (2017)
21. Kharazmi, E., Zhang, Z., Karniadakis, G.E.: Variational physics-informed neural networks for solving partial differential equations, (2019). arXiv preprint [arXiv:1912.00873](https://arxiv.org/abs/1912.00873)
22. Kharazmi, E., Zhang, Z., Karniadakis, G.E.: hp-vpinns: Variational physics-informed neural networks with domain decomposition. *Comput. Methods Appl. Mech. Eng.* **374**, 113547 (2021)
23. Lang, Q., Lu, F., Learning interaction kernels in mean-field equations of first-order systems of interacting particles. *SIAM J. Sci. Comput.* **44**(1), A260–A285 (2022)
24. Lang, Q., Lu, F., Identifiability of interaction kernels in mean-field equations of interacting particles. *Foundations of Data Science* **5**(4), 480–502 (2023)
25. Long, Z., Lu, Y., Ma, X., Dong, B. PDE-Net: Learning PDEs from Data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, page 9. PMLR, (2018)

26. Lu, F., Lang, Q., An, Q.: Data adaptive RKHS Tikhonov regularization for learning kernels in operators. *Proceedings of Mathematical and Scientific Machine Learning*, PMLR **190**, 158–172 (2022)
27. Lu, F., Maggioni, M., Tang, S.: Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *J. Mach. Learn. Res.* **22**(32), 1–67 (2021)
28. Lu, F., Maggioni, M., Tang, S.: Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Foundations of Computational Mathematics*, pages 1–55, (2021)
29. Fei, L., Zhong, M., Tang, S., Maggioni, M.: Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl. Acad. Sci. U.S.A.* **116**(29), 14424–14433 (2019)
30. Lu, Y., Li, X., Liu, C., Tang, Q., Wang, Y.: Learning generalized diffusions using an energetic variational approach, (2024)
31. Messenger, D.A., Bortz, D.M.: Bortz: Weak sindy for partial differential equations. *J. Comput. Phys.* **443**, 110525 (2021)
32. Messenger, D.A., Bortz, D.M.: Weak sindy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation* **19**(3), 1474–1497 (2021)
33. Messenger, D.A., Tran, A., Dukic, V., Bortz, D.M.: The weak form is stronger than you think, (2024). arXiv preprint [arXiv:2409.06751](https://arxiv.org/abs/2409.06751)
34. Øksendal, B.: Stochastic differential equations: an introduction with applications, 6th edn. Springer Science & Business Media, New York (2013)
35. Owahdi, H., Yoo, G.R.: Kernel flows: From learning kernels from data into the abyss. *J. Comput. Phys.* **389**, 22–47 (2019)
36. Ren, K., Zhang, L.: Data-driven joint inversions for PDE models, (2022). arXiv preprint [arXiv:2210.09228](https://arxiv.org/abs/2210.09228)
37. Schaeffer, H.: Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**(2197), 20160446 (2017)
38. Schafer, W.R.: Schafer: What is a Savitzky-Golay filter? *IEEE Signal Process. Mag.* **28**(4), 111–117 (2011)
39. Shampine, L.F.: Vectorized adaptive quadrature in matlab. *J. Comput. Appl. Math.* **211**(2), 131–140 (2008)
40. Sirignano, J., Spiliopoulos, K.: Dgm: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.* **375**, 1339–1364 (2018)
41. Song, W., Jiang, S., Camps-Valls, G., Mathew Williams, L., Zhang, M.R., Vereecken, H., He, L., Hu, X., Shi, L.: Towards data-driven discovery of governing equations in geosciences. *Communications Earth & Environment* **5**(1), 589 (2024)
42. Song, Y., Sohl-Dickstein, J., Diederik, P., Kingma, A., Kumar, S., Ermon, B., Poole: Score-based generative modeling through stochastic differential equations. In: *International Conference on Learning Representations*, (2021)
43. Stephany, R., Earls, C.: Weak-pde-learn: A weak form based approach to discovering pdes from noisy, limited data. *J. Comput. Phys.* **506**, 112950 (2024)
44. Villani, C.: Topics in optimal transportation, volume 58. American Mathematical Soc., (2003)
45. Wang, Y., Chen, J., Liu, C., Kang, L.: Particle-based energetic variational inference. *Stat. Comput.* **31**, 1–17 (2021)
46. Yang, L., Daskalakis, C., Karniadakis, G.E.: Generative ensemble regression: Learning particle dynamics from observations of ensembles with physics-informed deep generative models. *SIAM J. Sci. Comput.* **44**(1), B80–B99 (2022)
47. Yao, R., Chen, X., Yang, Y.: Mean-field nonparametric estimation of interacting particle systems. **99**, 6876897 (2022)
48. Zang, Y., Bao, G., Ye, X., Zhou, H.: Weak adversarial networks for high-dimensional partial differential equations. *J. Comput. Phys.* **411**, 109409 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.