Machine Learning (2025) 114:255
https://doi.org/10.1007/s10994-025-06908-9

Squared Wasserstein-2 loss functions for efficient learning of stochastic differential equations

Mingtao Xia[1] · Xiangting Li[2] · Qijing Shen[3] · Tom Chou[4]

Received: 3 July 2025 / Revised: 3 July 2025 / Accepted: 24 September 2025 /
Published online: 19 October 2025

Abstract

We provide an analysis of the squared Wasserstein-2 (W_2) distance between two probability distributions associated with two stochastic differential equations (SDEs). Based on this analysis, we propose using squared W_2 distance-based loss functions to train parametrized neural networks in order to reconstruct SDEs from noisy data. Specifically, we propose minimizing a time-decoupled squared W_2 distance loss function. To demonstrate the practicality of our Wasserstein distance-based loss functions, we performed numerical experiments that demonstrate the efficiency of our method in learning SDEs that arise across a number of applications.

Keywords Wasserstein distance · Stochastic differential equation · Inverse problem · Uncertainty quantification · Optimal transport

Mingtao Xia and Xiangting Li have contributed equally to this work.

Editor: Xiaolin Huang.

✉ Mingtao Xia
mxia4@uh.edu

✉ Xiangting Li
xiangting.li@ucla.edu

Qijing Shen
qijing.shen@ndm.oxford.edu

Tom Chou
tomchou@ucla.edu

[1] Department of Mathematics, University of Houston, Houston, TX 77204, USA

[2] Department of Computational Medicine, University of California Los Angeles, 621 Charles E. Young Dr. S., Los Angeles, CA 90095, USA

[3] Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK

[4] Department of Mathematics, University of California Los Angeles, 520 Portola Plaza, Los Angeles, CA 90095, USA

1 Introduction

Stochastic processes are mathematical models of random phenomena that evolve over time or space (Cinlar, 2011). Among stochastic processes, d-dimensional stochastic differential equations (SDE) of the form

$$\mathrm{d}\boldsymbol{X}(t) = f(\boldsymbol{X}(t), t)\mathrm{d}t + \sigma(\boldsymbol{X}(t), t)\mathrm{d}\boldsymbol{B}(t), \quad \boldsymbol{X}(t) \in \mathbb{R}^d, \quad t \in [0, T] \tag{1}$$

are widely used across different fields to model complex systems with continuous variables and noise. Here, $f = (f_1, ..., f_d) : \mathbb{R}^{d+1} \to \mathrm{R}^d$ and $\sigma = (\sigma_{i,j})_{i=1,\ldots,d,j=1,\ldots,s} : \mathrm{R}^{d+1} \to \mathrm{R}^{d\times s}$ denote deterministic and stochastic components of the SDE, while $\boldsymbol{B}(t)$ represents a s-dimensional standard Brownian motion. In applications such as computational fluid dynamics, cell biology, and genetics, the underlying dynamics are often unknown, partially observed, and subjected to noise. Consequently, it is vital to develop methods capable of learning the governing SDEs from limited data (Sullivan, 2015; Soize, 2017; Mathelin et al., 2005; Bressloff, 2014; Lin & Buchler, 2018). Traditional methods, such as the Kalman filtering (Welch et al., 1995; Welch, 2020) and Gaussian process regression (Liu et al., 2020; MacKay et al., 1998) often assume specific forms of noise. For example, (De Vecchi et al., 2016) uses polynomials to model f, σ, while (Pereira et al., 2010) assumes linear f and σ. If the forms of f and σ are known, then Bayesian methods are used to estimate the parameters therein (Gzyl et al., 2008). Such traditional methods may work well when prior information on the underlying SDE model is given. However, those methods may not be suitable for complex or nonlinear systems where noise affects the dynamics in a more complex manner and no prior information on f and σ in Eq. (1) is available.

Recent advancements leverage machine learning, specifically neural ordinary differential equations (NODEs) (Chen et al., 2018), to offer a more flexible approach to reconstructing SDEs in the form of neural SDEs (nSDEs) (Tzen and Raginsky, 2019; Tong et al., 2022; Jia & Benson, 2019). Previous attempts at using neural SDEs (nSDEs) have explored different loss functions for learning SDEs from data. For example, Tzen and Raginsky (2019) model the SDE as a continuum limit of latent deep Gaussian models and use a variational likelihood bound for training. Kidger et al. (2021) adopt Wasserstein generative adversarial networks (WGANs) that were proposed in Arjovsky et al. (2017) for reconstructing SDEs. Briol et al. (2019) uses a maximum mean discrepancy (MMD) loss and a generative model for training SDEs. Song et al. (2020) assumes that σ in Eq. (1) depends only on time and uses a score-based generative model for SDE reconstruction.

Despite promising recent advances, challenges remain, particularly in selecting optimal loss functions (Jia & Benson, 2019). The Wasserstein distance, a family of metrics that measures discrepancies between probability measures over a metric space, has emerged as a potential solution due to its robust properties (Villani et al., 2009; Oh et al., 2019; Zheng et al., 2020). Consequently, the Wasserstein distance, denoted as W, has gained wide use in statistics and machine learning. Key papers have delved into its analysis (Rüschendorf, 1985) and its utilization in reconstructing discrete-time stochastic processes (Bartl et al., 2021). In the context of SDEs, Bion-Nadal and Talay (2019) introduced a restricted Wasserstein-type distance, while Wang (2016) and Sanz-Serna and Zygalakis (2021) examined its application in ergodic SDEs, Levy processes, and Langevin equations, respectively. Calculating the W distance for multidimensional random variables is challenging; hence, approximations such as the sliced W distance and regularized W distance have emerged (Cuturi et al., 2019; Kolouri et al., 2018, 2019; Rowland et al., 2019; Frogner et al., 2015).

The aforementioned WGAN approach in Kidger et al. (2021) uses the first-order Wasserstein distance to indirectly learn SDEs via the Kantorovich-Rubinstein duality (Arjovsky et al., 2017). To the best of our knowledge, there has been no published work that directly analyzes the W distance and applies it to the learning of SDEs. In this paper, we introduce bounds on the second-order Wasserstein W_2 distance between two probability distributions over the continuous function space generated by solutions to two SDEs. Our results motivate the W_2 distance as the loss function to be used with parametrized neural networks for learning SDEs from time-series data containing intrinsic noise that results from Wiener processes. We test our approach on different examples to showcase its effectiveness.

2 Definitions and outline

We propose a squared W_2-distance-based loss function for training a neural-network-parametrized SDE model (Li et al., 2020) in order to reconstruct an SDE under the following setting. Let μ denote the probability distribution over the continuous function space $C([0,T];\mathbb{R}^d)$ generated by the solution $\boldsymbol{X}(t)$ to Eq. (1). In the following approximation to Eq. (1),

$$\mathrm{d}\hat{\boldsymbol{X}}(t)=\hat{f}(\boldsymbol{X}(t),t)\mathrm{d}t+\hat{\sigma}(\hat{\boldsymbol{X}}(t),t)\mathrm{d}\hat{\boldsymbol{B}}(t),\ \ t\in[0,T], \tag{2}$$

$\hat{\boldsymbol{B}}(t)$ is another s-dimensional standard Brownian motion independent of $\boldsymbol{B}(t)$ in Eq. (1), $\hat{f}=(\hat{f}_1,...,\hat{f}_d):\mathbb{R}^{d+1}\to\mathrm{R}^d$, and $\hat{\sigma}=(\hat{\sigma}_{i,j})_{i=1,...,d,j=1,...,s}:\mathrm{R}^{d+1}\to\mathrm{R}^{d\times s}$. The probability distribution over the continuous function space $C([0,T];\mathbb{R}^d)$ generated by the solution $\hat{\boldsymbol{X}}(t)$ to Eq. (2) will be denoted $\hat{\mu}$.

We shall follow the definition of the squared W_2-distance in Clement and Desch (2008) for two probability measures $\mu,\hat{\mu}$ associated with two continuous stochastic processes $\{\boldsymbol{X}(t)\}_{t\in[0,T]},\{\hat{\boldsymbol{X}}(t)\}_{t\in[0,T]}$.

Definition 1 For two d-dimensional continuous stochastic processes in the separable space $\big(C([0,T];\mathbb{R}^d),\|\cdot\|\big)$

$$\boldsymbol{X}(t)=\big(X^1(t),...,X^d(t)\big),\ \ \hat{\boldsymbol{X}}(t)=\big(\hat{X}^1(t),...,\hat{X}^d(t)\big),\ \ t\in[0,T], \tag{3}$$

with two associated probability distributions $\mu,\hat{\mu}$, the squared $W_2(\mu,\hat{\mu})$ distance between $\mu,\hat{\mu}$ is defined as

$$W_2^2(\mu,\hat{\mu})=\inf_{\pi(\mu,\hat{\mu})}\mathbb{E}_{(\boldsymbol{X},\hat{\boldsymbol{X}})\sim\pi(\mu,\hat{\mu})}\big[\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2\big]. \tag{4}$$

Throughout this paper, $\mathbb{E}$ refers to taking the expectation of a random variable, and $\mathbb{E}_{(\boldsymbol{X},\hat{\boldsymbol{X}})\sim\pi(\mu,\hat{\mu})}\big[\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2\big]$ refers to the expectation of the quantity $\big[\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2\big]$ when $(\boldsymbol{X},\hat{\boldsymbol{X}})$ obey the joint probability measure $\pi(\mu,\hat{\mu})$. The distance

$\|\boldsymbol{X}\| := \big(\int_0^T |X_i(t)|^2 \mathrm{d}t\big)^{\frac{1}{2}}$, where $|\cdot|$ is the l^2 norm of a vector. $\pi(\mu, \hat{\mu})$ iterates over all *coupled* distributions of $\boldsymbol{X}(t), \hat{\boldsymbol{X}}(t)$, defined by the condition

$$\begin{cases} \boldsymbol{P}_{\pi(\mu,\hat{\mu})}\left(A \times C([0,T];\mathbb{R}^d)\right) = \boldsymbol{P}_{\mu}(A), \\ \boldsymbol{P}_{\pi(\mu,\hat{\mu})}\left(C([0,T];\mathbb{R}^d) \times A\right) = \boldsymbol{P}_{\hat{\mu}}(A), \end{cases} \quad \forall A \in \mathcal{B}\Big(C([0,T];\mathbb{R}^d)\Big), \tag{5}$$

where $\mathcal{B}\Big(C([0,T];\mathbb{R}^d)\Big)$ denotes the Borel σ-algebra associated with the space of d-dimensional continuous functions $C([0,T];\mathbb{R}^d)$. Here, we assume that taking the expectation of the squares of the stochastic processes at a fixed time point is interchangeable with integration over time, i.e., for a stochastic process $\{\boldsymbol{X}(t)\}_{t=0}^T$,

$$\mathbb{E}\Big[\int_0^T |\boldsymbol{X}(t)|^2 \mathrm{d}t\Big] = \mathbb{E}\Big[\int_0^T |\boldsymbol{X}(t)|^2 \mathrm{d}t\Big] = \int_0^T \mathbb{E}\big[|\boldsymbol{X}(t)|^2\big]\mathrm{d}t. \tag{6}$$

Eq. (6) holds true for solutions to the SDE (1) under specific conditions, such as uniform bounds and Lipschitiz continuity on the coefficients f and σ which ensures a strong solution of the SDE. Detailed analysis on the interchangeability of taking the expectation and taking the integration w.r.t. time for stochastic processes are described by the stochastic Fubini theorem Jacod (2006); Choulli and Schweizer (2024).
The main contributions of our work are

1. Using Definition 1, we first derive in Sect. 3 an upper bound for the squared Wasserstein distance $W_2^2(\mu, \hat{\mu})$ between the probability measures associated with solutions to two 1D SDEs in terms of the errors in the reconstructed drift and diffusion functions, $f - \hat{f}$ and $\sigma - \hat{\sigma}$ in Eqs. (1) and (2). To be specific, we establish a W_2 distance upper bound which depends explicitly on the difference in the drift and diffusion functions $f - \hat{f}$ and $\sigma - \hat{\sigma}$ associated with using Eq. (2) to approximate Eq. (1).
2. In Sect. 4, we shall prove that the squared W_2 distance between the two SDEs, $W_2^2(\mu, \hat{\mu})$, can be accurately approximated by estimating the W_2 distance between their finite-dimensional projections. We also develop a time-decoupled squared Wasserstein-2 distance defined by

$$\tilde{W}_2^2(\mu, \hat{\mu}) := \int_0^T W_2^2(\mu(s), \hat{\mu}(s))\mathrm{d}s, \tag{7}$$

   which allows us to define a time-decoupled squared W_2-distance-based loss function for learning SDEs. Here, $\mu(s), \hat{\mu}(s)$ are the distributions on $\mathbb{R}^d$ generated by projection of the stochastic processes $\boldsymbol{X}, \hat{\boldsymbol{X}}$ at time s, respectively. We prove that the time-decoupled squared W_2 distance in Eq. (7) is well defined in Theorem 3, and that it inherits the upper bound of the squared Wasserstein distance $W_2^2(\mu, \hat{\mu})$ and could be evaluated using finite-time-point distributions of solutions to two SDEs. Specifically, if $X(t_i)$ follows the one-dimensional ($d = 1$) SDE Eq. (1), then for uniformly spaced time points $t_i = \frac{iT}{N},\ i = 0, ..., N$, our proposed time-decoupled squared W_2 loss function is simply

$$\Delta t \sum_{i=1}^{N-1} \int_0^1 \left(F_i^{-1}(s) - \hat{F}_i^{-1}(s)\right)^2 \mathrm{d}s, \tag{8}$$

where Δt is the timestep and F_i and $\hat{F}_i$ are the empirical cumulative distribution functions for $X(t_i)$ and $\hat{X}(t_i)$, respectively. This time-decoupled squared W_2-distance loss function will be explicitly expressed in Eq. (29).

3. Finally, we carry out numerical experiments to show that our squared W_2-distance-based SDE learning method performs better than recently developed machine-learning-based methods across many SDE reconstruction problems. Additional numerical experiments and sensitivity analysis are detailed in the Appendix.

3 Squared W_2 distance for learning SDEs

In this section, we prove the bounds for the squared W_2 distance of two probability measures associated with two SDEs. Specifically, we demonstrate that minimizing the squared W_2 distance is necessary for the reconstruction of f, σ in Eq. (1).

We shall first prove an upper bound for the W_2 distance between the probability measures μ and $\hat{\mu}$ associated with $X(t), \hat{X}(t)$, solutions to Eqs. (1) and (2), respectively.

Theorem 1 (The upper bound of the squared W_2 distance between distributions of solutions to two SDEs) We assume that $\{X(t)\}_{t\in[0,T]}, \{\hat{X}(t)\}_{t\in[0,T]}$ are solutions to Eqs. (1) and (2) (for $d=1$), respectively, and have the same distribution of initial conditions. Further requiring $f, \hat{f}, \sigma, \hat{\sigma}$ to be continuously differentiable, $\partial_x\sigma$ and $\partial_x\hat{\sigma}$ are uniformly bounded and

$$\begin{aligned}
W_2^2(\mu,\hat{\mu}) \leq & 3\int_0^T \mathbb{E}\Big[\int_0^t H^2(s,t)\mathrm{d}s\Big]\mathrm{d}t \times \mathbb{E}\Big[\int_0^T (f-\hat{f})^2(\tilde{X}(t),t)\mathrm{d}t)\Big] \\
& + 3\int_0^T \mathbb{E}\Big[\int_0^t H^2(s,t)\mathrm{d}s\Big]\mathrm{d}t \times \mathbb{E}\Big[\int_0^T \big(\partial_x\sigma(\eta_2(X(t),\tilde{X}(t)),t\big)^2(\sigma-\hat{\sigma})^2(\tilde{X}(t),t)\mathrm{d}t\Big] \\
& + 3\int_0^T \mathbb{E}\Big[\int_0^t H^4(s,t)\mathrm{d}s\Big]^{1/2}\mathrm{d}t \times \mathbb{E}\Big[\int_0^T (\sigma-\hat{\sigma})^4(\tilde{X}(t),t)\mathrm{d}t)\Big]^{1/2},
\end{aligned} \tag{9}$$

where $\tilde{X}(t)$ satisfies

$$\mathrm{d}\tilde{X}(t) = \hat{f}(\tilde{X}(t),t)\mathrm{d}t + \hat{\sigma}(\tilde{X}(t),t)\mathrm{d}B(t), \ \ \tilde{X}(0) = X(0). \tag{10}$$

In Eq. (9), $\eta_1, \eta_2 : \mathbb{R}^2 \to \mathbb{R}$ are two auxiliary functions such that

$$\begin{aligned}
f(X_1,t) - f(X_2,t) &= \partial_x f(\eta_1(X_1,X_2),t)(X_1 - X_2) \\
\sigma(X_1,t) - \sigma(X_2,t) &= \partial_x \sigma(\eta_2(X_1,X_2)t)(X_1 - X_2).
\end{aligned} \tag{11}$$

$$H(s,t) := \exp\left[\int_s^t h(X(r),\tilde{X}(r),r)\mathrm{d}r + \int_s^t \partial_x\sigma\big(\eta_2(X(r),\tilde{X}(r),r\big)\mathrm{d}B(r)\right], \tag{12}$$

with h defined as

$$h(X(r),\hat{X}(r),r):=\partial_x f\Big(\eta_1(X(r),\tilde{X}(r)),r\Big)-\Big(\partial_x\sigma\Big(\eta_2(X(r),\tilde{X}(r)),r\Big)\Big)^2. \tag{13}$$

The proof to Theorem 1 and its generalizations to higher dimensional stochastic dynamics under some specific assumptions are given in Appendix A. Theorem 1 indicates that as long as $\mathbb{E}\big[\int_0^t H^4(s,t)\mathrm{d}s\big]$ is uniformly bounded for all $t\in[0,T]$, the upper bound for $W_2(\mu,\hat{\mu})\to 0$ when $\hat{f}-f\to 0$ and $\hat{\sigma}-\sigma\to 0$ uniformly in $\mathbb{R}\times[0,T]$. Specifically, if $f=\hat{f},\sigma=\hat{\sigma}$, then the RHS Eq. (9) is 0. This means that minimizing $W_2^2(\mu,\hat{\mu})$ is necessary for generating small errors $\hat{f}-f,\hat{\sigma}-\sigma$ and for accurately approximating both *f* and σ. Thus, one can consider using the squared W_2 distance as an effective loss function to minimize when learning SDEs from data. MSE-based loss functions (defined in Appendix E) suppress noise while the Kullback-Liebler (KL) divergence may not be finite, precluding resolution of $X(t)$ and $\hat{X}(t)$ even if $\hat{f}$ approximates f and $\hat{\sigma}$ approximates σ. Detailed discussions on the limitations of MSE and KL divergence in SDE reconstruction can be found in Appendix B.

Remark. A generalized version of Theorem 1 with relaxed conditions for the upper bound of the squared W_2 distance between two multidimensional pure-diffusion and jump-diffusion processes is given in subsequent work Xia et al. (2024). Consider $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ describing general *d*-dimensional SDEs Eqs. (1) and (2). It is nontrivial to show whether or not the squared W_2 distance between two multidimensional pure-diffusion processes or two jump-diffusion processes is an upper bound for the errors $\hat{f}-f,\hat{\sigma}-\sigma$. If it is, minimizing $W_2^2(\mu,\hat{\mu})$ is sufficient for reconstructing of f,σ using $\hat{f},\hat{\sigma}$. However, in Xia et al. (2024), some preliminary results on how the squared W_2 distance might serve as an upper bound for the errors $\hat{f}-f,\hat{\sigma}-\sigma$ in 1D jump-diffusion processes are given. Theorem 2.1 in Xia et al. (2024) indicates that as long as $f,\sigma,\hat{f},\hat{\sigma}$ are continuously differentiable and uniformly Lipschitz continuous, then

$$W_2(\mu,\hat{\mu})\leq\sqrt{T\,\mathbb{E}[H(T)|\boldsymbol{X}(0)]}\times\exp\left(CT\right), \tag{14}$$

where *C* is a constant depending on $f,\hat{f},\sigma,\hat{\sigma}$, $\mu,\hat{\mu}$ are the probability distribution over the continuous function space $C([0,T];\mathbb{R}^d)$ generated by the solutions $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$, respectively. Furthermore, in Eq. (14),

$$\begin{aligned}H(t):=\;&\mathbb{E}\left[\sum_{i=1}^{d}\int_0^t\Big(f_i(X(s),s)-\hat{f}_i(X(s),s)\Big)^2\,\mathrm{d}s\right]\\&+\mathbb{E}\left[\sum_{i=1}^{d}\int_0^t\sum_{j=1}^{s}\big(\sigma_{i,j}(X(s),s)-\hat{\sigma}_{i,j}(X(s),s)\big)^2\,\mathrm{d}s\right].\end{aligned} \tag{15}$$

4 Finite-dimensional and time-decoupled squared W_2 loss functions

From Theorem 1 in Sect. 3, in order to have small errors in the drift and diffusion terms $f-\hat{f}$ and $\sigma-\hat{\sigma}$, a small $W_2(\mu,\hat{\mu})$ is necessary. However, $W_2(\mu,\hat{\mu})$ cannot be directly used as a loss function to minimize since we cannot directly evaluate the integration in time in Eq. (4). In this section, we shall provide a way to estimate the $W_2(\mu,\hat{\mu})$ distance by using finite-dimensional projections, leading to squared W_2-distance-based loss functions for minimization.

Consider the two general d-dimensional SDEs defined in Eqs. (1) and (2). Usually, we only have observations of trajectories of $\{\boldsymbol{X}(t)\}_{t\in[0,T]}$ and $\{\hat{\boldsymbol{X}}(t)\}_{t\in[0,T]}$ over discrete time points. We assume that $\boldsymbol{X}(t), \hat{\boldsymbol{X}}(t)$ solve the two SDEs described by Eqs. (1) and (2) and provide an estimate of the W_2 between of the probability measures $\mu, \hat{\mu}$ associated with $\{\boldsymbol{X}(t)\}_{t\in[0,T]}$ and $\{\hat{\boldsymbol{X}}(t)\}_{t\in[0,T]}$ using their finite-dimensional projections. We let $0 = t_0 < t_1 < ... < t_N = T, t_i = i\Delta t, \Delta t := \frac{T}{N}$ be a uniform mesh in time and define the following projection operator $\boldsymbol{I}_N$

$$\boldsymbol{X}_N(t) := \boldsymbol{I}_N\boldsymbol{X}(t) = \begin{cases}\boldsymbol{X}(t_i), t\in[t_i, t_{i+1}), i < N-1,\\ \boldsymbol{X}(t_i), t\in[t_i, t_{i+1}], i = N-1.\end{cases} \tag{16}$$

As in the previous case, we require $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ to be continuous. Note that the projected process is no longer continuous. Thus, we define a new space $\tilde{\Omega}_N$ containing all continuous and piecewise constant functions; naturally, $\mu, \hat{\mu}$ are allowed to be defined on $\tilde{\Omega}_N$. Distributions over $\tilde{\Omega}_N$ generated by $\{\boldsymbol{X}_N(t)\}_{t\in[0,T]}$,$\{\hat{\boldsymbol{X}}_N(t)\}_{t\in[0,T]}$ in Eq. (16) is denoted by μ_N and $\hat{\mu}_N$, respectively. We will prove the following theorem for estimating $W_2(\mu,\hat{\mu})$ by $W_2\left(\mu_N, \hat{\mu}_N\right)$.

Theorem 2 [Finite-time-point approximation of the squared W_2 distance] Suppose $\{\boldsymbol{X}(t)\}_{t\in[0,T]}$ and $\{\hat{\boldsymbol{X}}(t)\}_{t\in[0,T]}$ are both continuous-time continuous-space stochastic processes in $\mathbb{R}^d$ and $\mu, \hat{\mu}$ are their associated probability measures, then $W_2(\mu,\hat{\mu})$ can be bounded by their finite-dimensional projections

$$W_2(\mu_N,\hat{\mu}_N) - W_2(\mu,\mu_N) - W_2(\hat{\mu},\hat{\mu}_N) \le W_2(\mu,\hat{\mu}) \le W_2(\mu_N,\hat{\mu}_N) + W_2(\mu,\mu_N) + W_2(\hat{\mu},\hat{\mu}_N) \tag{17}$$

where $\mu_N, \hat{\mu}_N$ are the probability distributions associated with the two stochastic processes $\{\boldsymbol{X}_N(t)\}_{t\in[0,T]}$ and $\{\hat{\boldsymbol{X}}_N(t)\}_{t\in[0,T]}$ defined in Eq. (16). Specifically, if $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ solve Eqs. (1) and (2), and if

$$\begin{aligned}
F :=& \mathbb{E}_{\boldsymbol{X}\sim\mu}\Big[\int_0^T \sum_{i=1}^d f_i^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big] < \infty,\\
\hat{F} :=& \mathbb{E}_{\hat{\boldsymbol{X}}\sim\hat{\mu}}\Big[\int_0^T \sum_{i=1}^d \hat{f}_i^2(\hat{\boldsymbol{X}}(t), t)\mathrm{d}t\Big] < \infty,\\
\Sigma :=& \mathbb{E}_{\boldsymbol{X}\sim\mu}\Big[\int_0^T \sum_{\ell=1}^d\sum_{j=1}^s \sigma_{i,j}^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big] < \infty,\\
\hat{\Sigma} :=& \mathbb{E}_{\hat{\boldsymbol{X}}\sim\hat{\mu}}\Big[\int_0^T \sum_{\ell=1}^d\sum_{j=1}^s \hat{\sigma}_{i,j}^2(\hat{\boldsymbol{X}}(t), t)\mathrm{d}t\Big] < \infty,
\end{aligned} \tag{18}$$

then we obtain the following bound

$$\begin{aligned} W_2(\mu_N, \hat{\mu}_N) - \sqrt{(s+1)\Delta t}\left(\sqrt{F\Delta t + \Sigma} + \sqrt{\hat{F}\Delta t + \hat{\Sigma}}\right) &\le W_2(\mu, \hat{\mu}) \\ \le W_2(\mu_N, \hat{\mu}_N) + \sqrt{(s+1)\Delta t}&\left(\sqrt{F\Delta t + \Sigma} + \sqrt{\hat{F}\Delta t + \hat{\Sigma}}\right). \end{aligned} \tag{19}$$

The proof to Theorem 2 relies on the triangular inequality of the Wasserstein distance and the Itô isometry; it is provided in Appendix C. Theorem 2 gives bounds for approximating the W_2 distance between the distributions of $\{\boldsymbol{X}(t)\}_{t\in[0,T]}, \{\hat{\boldsymbol{X}}(t)\}_{t\in[0,T]}$ by the W_2 distance between the distributions of their finite-time-point projections $\{\boldsymbol{X}_N(t)\}_{t\in[0,T]}, \{\hat{\boldsymbol{X}}_N(t)\}_{t\in[0,T]}$. Specifically, if $\boldsymbol{X}(t), \hat{\boldsymbol{X}}(t)$ are solutions to Eqs. (1) and (2), then as the timestep $\Delta t \to 0$, $W_2(\mu_N, \hat{\mu}_N) \to W_2(\mu, \hat{\mu})$. Theorem 2 indicates that we can use $W_2^2(\mu_N, \hat{\mu}_N)$, which approximates $W_2^2(\mu, \hat{\mu})$ when $\Delta t \to 0$, as a loss function. Furthermore,

$$W_2^2(\mu_N, \hat{\mu}_N) = \inf_{\pi(\mu_N, \hat{\mu}_N)} \sum_{i=1}^{N-1} \mathbb{E}_{(\boldsymbol{X}_N, \hat{\boldsymbol{X}}_N)\sim\pi(\mu_N, \hat{\mu}_N)} \left[\left|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\right|_2^2\right] \Delta t. \tag{20}$$

here, $\pi(\mu_N, \hat{\mu}_N)$ iterates over coupled distributions of $\{\boldsymbol{X}_N(t)\}_{t\in[0,T]}, \{\hat{\boldsymbol{X}}_N(t)\}_{t\in[0,T]}$, whose marginal distributions coincide with μ_N and $\hat{\mu}_N$. $|\cdot|_2$ denotes the ℓ^2 norm of a vector. Note that μ_N is fully characterized by values of $\{\boldsymbol{X}(t)\}_{t\in[0,T]}$ at the discrete time points t_i.

Remark. The *d*-dimensional SDEs in Eqs. (1) and (2) can be solved numerically. Solutions to the two SDEs can be approximated by strong, order γ Itô-Taylor solutions; we will denote these by $\{\boldsymbol{X}_\delta(t)\}_{t\in[0,T]}$ and $\{\hat{\boldsymbol{X}}_\delta(t)\}_{t\in[0,T]}$, along with their associated probability distributions denoted by μ_δ and $\hat{\mu}_\delta$. Here, δ denotes a uniform time step used in the numerical scheme, which can be different from Δt in Eq. (20). For simplicity, we can assume that Δt is an integer multiple of δ and that all coefficients involved in the order γ Itô-Taylor scheme satisfy the conditions prescribed in (Kloeden and Platen (1992), Theorem 10.6.3). Then, using Theorem 10.6.3 in Kloeden and Platen (1992), we have the following result which takes into account the time discretization error of the numerical SDE scheme.

Corollary 1 Suppose $\boldsymbol{X}_\delta(t)$ and $\hat{\boldsymbol{X}}_\delta(t)$ (δ denotes a uniform time step) are numerical solutions of order γ strong Ito-Taylor approximates to Eqs. (1) and (2) with all involved coefficients satisfying the conditions specified in (Kloeden and Platen (1992), Theorem 10.6.3). We denote $\mu_{\delta,N}, \hat{\mu}_{\delta,N}$ to be the distributions of $\boldsymbol{I}_N\boldsymbol{X}_\delta(t)$ and $\boldsymbol{I}_N\hat{\boldsymbol{X}}_\delta(t)$, respectively. Suppose $\boldsymbol{X}_\delta(0) = \hat{\boldsymbol{X}}_\delta(0) = \boldsymbol{X}(0)$, then the following inequality holds:

$$\begin{aligned} W_2(\mu_{\delta,N}, \hat{\mu}_{\delta,N}) - W_2(\mu, \mu_N) - W_2(\hat{\mu}, \hat{\mu}_N) - K(1+|\boldsymbol{X}(0)|^2)^{\frac{1}{2}}\delta^\gamma\sqrt{T} &\le W_2(\mu, \hat{\mu}) \\ \le W_2(\mu_{\delta,N}, \hat{\mu}_{\delta,N}) + W_2(\mu, \mu_N) + W_2(\hat{\mu}, \hat{\mu}_N) + K(1+|\boldsymbol{X}(0)|^2)^{\frac{1}{2}}\delta^\gamma\sqrt{T}&, \end{aligned} \tag{21}$$

where K is a constant that does not depend on δ.

Proof The proof of Corollary 1 is a straightforward application of Theorem 2 and (Kloeden and Platen (1992), Theorem 10.6.3). Notice that

$$W_2(\mu_N, \hat{\mu}_N) \ge W_2(\mu_{\delta,N}, \hat{\mu}_{\delta,N}) - W_2(\mu_{\delta,N}, \mu_N) - W_2(\hat{\mu}_{\delta,N}, \hat{\mu}_N), \tag{22}$$

and

$$W_2(\mu_N, \hat{\mu}_N) \leq W_2(\mu_{\delta,N}, \hat{\mu}_{\delta,N}) + W_2(\mu_{\delta,N}, \mu_N) + W_2(\hat{\mu}_{\delta,N}, \hat{\mu}_N). \tag{23}$$

Furthermore,

$$W_2^2(\mu_{\delta,N}, \mu_\delta) = \inf_{\pi(\mu_N, \hat{\mu}_N)} \sum_{i=1}^{N-1} \mathbb{E}_{(\boldsymbol{X}_N, \hat{\boldsymbol{X}}_{\delta,N}) \sim \pi(\mu_N, \mu_{\delta,N})} \Big[\big| \boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}_\delta(t_i) \big|_2^2 \Big] \Delta t. \tag{24}$$

here, $\pi(\mu_N, \mu_{\delta,N})$ iterates over coupled distributions of $\boldsymbol{X}_N(t), \boldsymbol{X}_{\delta,N}(t)$, whose marginal distributions coincide with μ_N and $\mu_{\delta,N}$. We take a special coupling such that

$$\begin{aligned} \boldsymbol{X}(t) &= \int_0^t \boldsymbol{f}(\boldsymbol{X}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{X}(s), s)\mathrm{d}\boldsymbol{B}_s + \boldsymbol{X}(0),\ t \in [0, T], \\ \boldsymbol{X}_\delta\big((i+1)\delta\big) &= \sum_{\alpha \in A_\gamma} f_\alpha\big(i\delta, \boldsymbol{X}(i\delta)\big) I_\alpha, \end{aligned} \tag{25}$$

where α and f_α are the indices and coefficients in (Kloeden and Platen (1992), Theorem 10.6.3) and I_α is the multiple Itô integral for the index α associated with $\boldsymbol{B}_t$. Using (Kloeden and Platen (1992), Theorem 10.6.3), there exists a δ-independent constant K' such that

$$\mathbb{E}\Big(\sup_{0 \leq t \leq T} |\boldsymbol{X}(t) - \boldsymbol{X}_\delta(t)|^2 \Big) \leq K'(1 + |\boldsymbol{X}(0)|^2)\delta^{2\gamma}. \tag{26}$$

Then, from Eq. (24), we conclude that

$$W_2^2(\mu_{\delta,N}, \mu_\delta) \leq N\Delta t K'(1 + |\boldsymbol{X}(0)|^2)\delta^{2\gamma} = TK'(1 + |\boldsymbol{X}(0)|^2)\delta^{2\gamma}. \tag{27}$$

Similarly,

$$W_2^2(\hat{\mu}_{\delta,N}, \hat{\mu}_\delta) \leq N\Delta t K'(1 + |\boldsymbol{X}(0)|^2)\delta^{2\gamma} = TK'(1 + |\boldsymbol{X}(0)|^2)\delta^{2\gamma}. \tag{28}$$

Defining $K := 2\sqrt{K'}$, the inequality (21) is proved. □

For a d-dimensional SDE, the trajectories at discrete time points $\{\boldsymbol{X}(t_i)\}_{i=1}^{N-1}$ is $d \times (N-1)$ dimensional. In Fournier and Guillin (2015), the error bound for $|W_2^2(\mu_N, \hat{\mu}_N) - W_2^2(\mu_N^{\mathrm{e}}, \hat{\mu}_N^{\mathrm{e}})|$, where $\mu_N^{\mathrm{e}}, \hat{\mu}_N^{\mathrm{e}}$ are the finite-sample empirical distributions of $\{\boldsymbol{X}(t_i)\}_{i=1}^{N-1}$ and $\{\hat{\boldsymbol{X}}(t_i)\}_{i=1}^{N-1}$, will increase as the dimensionality $d \times (N-1)$ becomes large. Alternatively, we can disregard the temporal correlations of values at different times and relax the constraint on the coupling $\pi(\mu_N, \hat{\mu}_N)$ in to minimize the Wasserstein distance between the marginal distribution of $\{\boldsymbol{X}(t_i)\}_{i=1}^{N-1}$ and the marginal distribution of $\{\hat{\boldsymbol{X}}(t_i)\}_{i=1}^{N-1}$, as was done in Chewi et al. (2021). To be more specific, we minimize individual terms in the sum with respect to the coupling π_i of the distributions of $\boldsymbol{X}(t_i)$ and $\hat{\boldsymbol{X}}(t_i)$ and define a heuristic loss function

$$\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\Delta t=\sum_{i=1}^{N-1}W_2^2(\mu_N(t_i),\hat{\mu}_N(t_i))\Delta t \tag{29}$$

where $\mu_N(t)$ and $\hat{\mu}_N(t)$ are the probability distributions of $\boldsymbol{X}(t)$ and $\boldsymbol{X}(t)$ at time t, respectively. Note that

$$\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\Delta t\le W_2^2(\mu_N,\hat{\mu}_N) \tag{30}$$

because the marginal distributions of $\pi(\mu_N,\hat{\mu}_N)$ coincide with μ_N and $\hat{\mu_N}$. Since the marginal distributions of μ_N and $\hat{\mu}_N$ at t_i are $\mu_N(t_i)$ and $\hat{\mu}_N(t_i)$, respectively, we have

$$\begin{aligned}&\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\Delta t\\&\qquad\le\inf\nolimits_{\pi(\mu_N,\hat{\mu}_N)}\sum_{i=1}^{N-1}\mathbb{E}_{(\boldsymbol{X}_N,\hat{\boldsymbol{X}}_N)\sim\pi(\mu_N,\hat{\mu}_N)}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\Delta t.\end{aligned} \tag{31}$$

The dimensionality of $\boldsymbol{X}(t_i)$ and $\hat{\boldsymbol{X}}(t_i)$ is d, which is much smaller than $(N-1)d$ for large N. We denote $\mu_N^{\mathrm{e}}(t_i)$ and $\hat{\mu}_N^{\mathrm{e}}(t_i)$ to be the finite-sample empirical distributions of $\boldsymbol{X}(t_i)$ and $\hat{\boldsymbol{X}}(t_i)$, respectively. Since the error of estimating the W_2 distance using empirical distributions of a random variable increases with the random variable's dimensionality Fournier and Guillin (2015), the error $\big|\sum_{i=1}^{N-1}W_2^2(\mu_N(t_i),\hat{\mu}_N(t_i))-\sum_{i=1}^{N-1}W_2^2(\mu_N^{\mathrm{e}}(t_i),\hat{\mu}_N^{\mathrm{e}}(t_i))\big|$ can be smaller than the error $\big|W_2^2(\mu_N,\hat{\mu}_N)-W_2^2(\mu_N^{\mathrm{e}},\hat{\mu}_N^{\mathrm{e}})\big|$. Compared to Eq. (20), the time-decoupled squared W_2 distance Eq. (29) can be better approximated using finite-sample empirical distributions.

Note that

$$\sum_{i=1}^{N-1}W_2^2(\mu_N(t_i),\hat{\mu}_N(t_i))\Delta t\le W_2^2(\mu_N,\hat{\mu}_N). \tag{32}$$

Thus, from Theorems 1 and 2, minimizing Eq. (29) when $N\to\infty$ is also necessary to achieve small $f-\hat{f}$ and $\sigma-\hat{\sigma}$ when the SDE is univariate. Let $\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i$ be the two probability distributions on the space of continuous functions associated with $\boldsymbol{X}(t),t\in[t_i,t_{i+1})$ and $\hat{\boldsymbol{X}}(t),t\in[t_i,t_{i+1})$, respectively. We can then show that Eq. (29) is an approximation to the partially time-decoupled summation of squared W_2 distances $\sum_{i=1}^{N-1}W_2^2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i)$ as $N\to\infty$. Additionally, we can prove the following theorem that indicates Eq. (29) approximates a time-decoupled squared Wasserstein distance Eq. (7) in the $N\to\infty$ limit.

Theorem 3 [Well-posedness of our proposed time-decoupled squared W_2 distance Eq. (7)] We assume the conditions in Theorem 2 hold and for any $0<t<t'<T$, as $t'-t\to0$, the following conditions are satisfied

$$\begin{aligned}&\mathbb{E}\Big[\int_t^{t'}\sum_{i=1}^{d} f_i^2(\boldsymbol{X}(t),t)\mathrm{d}t\Big],\ \mathbb{E}\Big[\int_t^{t'}\sum_{i=1}^{d} \hat{f}_i^2(\hat{\boldsymbol{X}}(t),t)\mathrm{d}t\Big]\to 0,\\ &\mathbb{E}\Big[\int_t^{t'}\sum_{i=1}^{d}\sum_{j=1}^{s}\sigma_{i,j}^2(\boldsymbol{X}(t),t)\mathrm{d}t\Big],\ \mathbb{E}\Big[\int_t^{t'}\sum_{i=1}^{d}\sum_{j=1}^{s}\hat{\sigma}_{i,j}^2(\hat{\boldsymbol{X}}(t),t)\mathrm{d}t\Big]\to 0.\end{aligned} \tag{33}$$

Then,

$$\lim_{N\to\infty}\Big(\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t-\sum_{i=1}^{N-1}W_2^2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i)\Big)=0. \tag{34}$$

Furthermore, the limit

$$\lim_{N\to\infty}\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t=\lim_{N\to\infty}\sum_{i=1}^{N-1}W_2^2\big(\mu(t_i),\hat{\mu}(t_i)\big)\Delta t \tag{35}$$

exists.

The proof of Theorem 3 will use the result of Theorem 2 and is given in Appendix D. Specifically, for each N,

$$\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t\le W_2^2(\mu_N,\hat{\mu}_N), \tag{36}$$

so we conclude that

$$\lim_{N\to\infty}\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t\le\lim_{N\to\infty}W_2^2(\mu_N,\hat{\mu}_N)=W_2^2(\mu,\hat{\mu}). \tag{37}$$

We denote

$$\tilde{W}_2^2(\mu,\hat{\mu}):=\int_0^T W_2^2\big(\mu(t),\hat{\mu}(t)\big)\mathrm{d}t=\lim_{N\to\infty}\sum_{i=1}^{N-1}W_2^2\big(\mu(t_i^1),\hat{\mu}(t_i^1)\big)(t_i^1-t_{i-1}^1) \tag{38}$$

as the *time-decoupled squared Wasserstein distance*. From Eq. (37), we can deduce that

$$\tilde{W}_2^2(\mu,\hat{\mu})\le W_2^2(\mu,\hat{\mu}). \tag{39}$$

Therefore, the upper bound of $W_2^2(\mu,\hat{\mu})$ in Theorem 1 is also an upper bound of $\tilde{W}_2^2(\mu,\hat{\mu})$, i.e., to reconstruct a 1D SDE by minimizing $\tilde{W}_2^2(\mu,\hat{\mu})$, it is necessary that $f-\hat{f}$ and $\sigma-\hat{\sigma}$ are small. From Theorem 3, minimizing the finite-time-point time-decoupled loss function defined in Eq. (29), which approximates $\tilde{W}_2^2(\mu, \hat{\mu})$ when Δt is small, is needed for minimizing $f - \hat{f}$ and $\sigma - \hat{\sigma}$.

Remark. If we replace $\boldsymbol{X}(t_i)$ and $\hat{\boldsymbol{X}}(t_i)$ in Eq. (35) with $\boldsymbol{X}_\delta(t_i)$ and $\hat{\boldsymbol{X}}_\delta(t_i)$, the order γ strong numerical solutions to Eqs. (1) and (2), and assuming the conditions in Corollary 1 hold,

$$\lim_{N\to\infty}\Big(\sum_{i=1}^{N-1} W_2^2\big(\mu_\delta(t_i), \hat{\mu}_\delta(t_i)\big)\Delta t - \sum_{i=1}^{N-1} W_2^2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_i)\Big) = 0, \tag{40}$$

where $\mu_\delta(t), \hat{\mu}_\delta(t)$ are the probability distributions of $\boldsymbol{X}_\delta(t)$ and $\hat{\boldsymbol{X}}_\delta(t)$ at time *t*, respectively. This arises because

$$\begin{aligned}&\Big|\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]^{1/2} - W_2\big(\mu_\delta(t_i), \hat{\mu}_\delta(t_i)\big)\Big|\\ &\qquad \le W_2(\mu(t_i), \mu_\delta(t_i)) + W_2(\hat{\mu}(t_i), \hat{\mu}_\delta(t_i)) \le K(1 + |\boldsymbol{X}(0)|^2)^{1/2}\delta^\gamma.\end{aligned} \tag{41}$$

Thus, since we assumed that Δt is an integer multiple of δ and thus $\Delta t \ge \delta$, we have

$$\begin{aligned}&\lim_{N\to\infty}\Big(\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t - \sum_{i=1}^{N-1} W_2^2\big(\mu_\delta(t_i), \hat{\mu}_\delta(t_i)\big)\Delta t\Big)\\ &\le \lim_{N\to\infty} 4\max_i\inf_{\pi_i}\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]^{1/2} K\delta^\gamma T(1 + |\boldsymbol{X}(0)|^2)^{1/2} + 4K^2\delta^{2\gamma}T(1 + |\boldsymbol{X}(0)|^2) = 0.\end{aligned} \tag{42}$$

Specifically, if $X(t), \hat{X}(t)$ are solutions to the univariate SDEs Eq. (1) and Eq. (2), then Eq. (29) reduces to Eq. (8), which can be directly calculated. In Example 3, Example 4, and Appendix I, we shall compare use of the two different squared W_2 distance loss functions Eqs. (20) and (29). From our preliminary numerical results, using Eq. (29) is more efficient than using Eq. (20) and yields reconstructed SDEs that are more accurate.

5 Numerical experiments

We carry out experiments to investigate the efficiency of our proposed squared W_2 loss function (Eq. (29)) by comparing it to other methods and loss functions. Our approach is tested on the reconstruction of several representative SDEs in Examples 1– 4.

In all experiments, we use two neural networks to parameterize $\hat{f} := \hat{f}(X, t; \Theta_1), \hat{\sigma} := \hat{\sigma}(X, t; \Theta_2)$ in Eq. (2) for the purpose of learning f, σ in Eq. (1) by the estimates $\hat{f} \approx f, \hat{\sigma} \approx \sigma$. Θ_1, Θ_2 are the parameter sets in the two neural networks for parameterizing $\hat{f} = \hat{f}_{\Theta_1}, \hat{\sigma} = \hat{\sigma}_{\Theta_2}$. We use the `sdeint` function in the `torchsde` Python package in Li et al. (2020) to numerically integrate SDEs. Details of the training hyperparameter setting for all examples are given in Table 1. A pseudocode for using the time-decoupled squared W_2 loss function Eq. (29) to train the neural networks $\hat{f}(X, t; \Theta_1)$ and $\hat{\sigma}(X, t; \Theta_2)$ is given in Algorithm 1. All experiments were carried out using Python 3.11 on a desktop with a 24-core Intel® i9-13900KF CPU. Default hyperparameters and training settings for each example are listed in Table 1, and the default Euler-Maruyama

Table 1 Training settings for each example

Loss	Example 1	Example 2	Example 3	Example 4	Example 5
Gradient descent method	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	0.001	0.002	0.002	0.0005	0.002
Weight decay	0.005	0.005	0.005	0.005	0.005
Number of epochs	1000	2000	2000	2000	500
Number of samples	100	200	256	200	100
Hidden layers in Θ_1	2	1	1	1	1
Neurons in each layer in Θ_1	32	32	32	32	150
Hidden layers in Θ_2	2	1	1	1	1
Activation function	tanh	ReLu	ReLu	ReLu	ReLu
Neurons in each layer in Θ_2	32	32	32	32	150
Δt	0.1	0.05	1	0.02	0.5

scheme (corresponding to the order $\gamma = \frac{1}{2}$ strong Ito Taylor expansion in Corollary 1) in the `torchsde` package is used for numerically solving SDEs in all numerical examples.

```
Given $M$ observed trajectories $\{X_i(t_j), t_j = j\Delta t, j = 1, ..., N\}_{i=1}^{M}$, and the maximal iteration $i_{\max}$
Initialize the two neural networks: $\hat{f}(X, t; \Theta_1)$ and $\hat{\sigma}(X, t; \Theta_2)$ in Eq. (2)
Generate $M$ trajectories from the approximate SDE Eq. (2) using the torchsde package
while $\sum_{i=1}^{N-1} W_2^2\big(\mu_N(t_i), \hat{\mu}_N(t_i)\big)\Delta t > \epsilon$ && $i < i_{\max}$ do
    Perform gradient descent to minimize the loss function $W_2^2(\mu_N, \hat{\mu}_N)$ and update the parameters in $\Theta_1, \Theta_2$ in $\hat{f}(X, t; \Theta_1)$ and $\hat{\sigma}(X, t; \Theta_2)$
    Generate $M$ trajectories from the approximate SDE Eq. (2) with the updated $\hat{f}(X, t; \Theta_1)$ and $\hat{\sigma}(X, t; \Theta_2)$ using the torchsde package
end while
return The trained approximate drift function $\hat{f}(X, t; \Theta_1)$ and diffusion function $\hat{\sigma}(X, t; \Theta_2)$
```

Algorithm 1 The pseudocode of minimizing the squared W_2 loss function to train a neural SDE. (The time-decoupled squared W_2 loss in the **while** loop can be replaced with other loss functions)

First, we compare our proposed squared W_2-distance-based loss (Eq. (29)) with several traditional statistical methods for SDE learning or reconstruction.

Example 1 We reconstruct a nonlinear SDE of the form

$$\mathrm{d}X(t) = \Big(\tfrac{1}{2} - \cos X(t)\Big)\mathrm{d}t + \sigma \mathrm{d}B(t), \quad t \in [0, 20], \tag{43}$$

which defines a Brownian process in a potential of the form $U(x) = \frac{x}{2} - \sin x$. In the absence of noise, there are infinitely many stable equilibrium points $x_k = \frac{5\pi}{3} + 2\pi k, k \in \mathbb{Z}$. When noise $\sigma \mathrm{d}B(t)$ is added, trajectories tend to saturate around those equilibrium points but jumping from one equilibrium point to another is possible. We set $\sigma \equiv 1$. We use the MSE, the mean^2+variance, the maximum-log-likelihood, and the proposed finite-time-point time-decoupled squared W_2 distance Eq. (29) as loss functions to reconstruct Eq. (43). For all loss functions, we use the same neural network hyperparameters. Definitions of all loss functions and training details are provided in Appendix E. As detailed in Table 1, neural networks with the same number of hidden layers and neurons in each layer are used for each loss function. Using the initial condition $X(0) = 0$, the sampled ground-truth and reconstructed trajectories are shown in Fig. 1.

Figure 1a shows the distributions of 100 trajectories with most of them concentrated around two attractors (local minima $x = -\frac{\pi}{3}, \frac{5\pi}{3}$ of the potential $U(x)$). Figure 1b shows that using MSE gives almost deterministic trajectories and fails to reconstruct the noise. From 1c, we see that the mean^2+variance loss fails to reconstruct the two local equilibria because cannot sufficiently resolve the shape of the trajectory distribution at any fixed timepoint. Figure 1d shows that when using our proposed finite-time-point time-decoupled squared W_2 loss Eq. (29), the trajectories of the reconstructed SDE can successfully learn the two-attractor feature and potentially the distribution of trajectories. The reason why the reconstructed trajectories of the W_2 distance cannot recover the third stable equilibrium at $x = \frac{11\pi}{3}$ is because the data is sparse near it. From 1e, we see that the max-log-likelihood loss performs the worst as it yields almost the same curves for all realizations.

In the next example, we show how using our finite-time-point time-decoupled squared W_2 distance loss function Eq. (29) can lead to efficient reconstruction of f and σ. We shall use the mean relative L^2 error

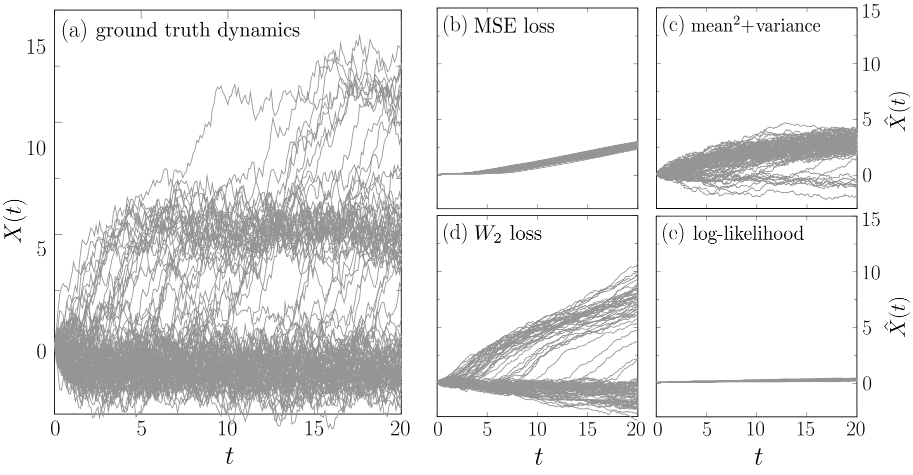

Fig. 1 **a** Ground-truth trajectories. **b** Reconstructed trajectories from nSDE using MSE loss. **c** Reconstructed trajectories from nSDE using mean^2+variance loss. **d** Reconstructed trajectories from nSDE using the finite-time-point time-decoupled W_2 loss. **e** Reconstructed trajectories from nSDE using a max-log-likelihood loss yields the worst approximation

$$\Big(\sum_{i=0}^{T}\frac{\sum_{j=1}^{N}\|f(x_j(t_i),t_i)-\hat{f}(x_j(t_i),t_i)\|^2}{(T+1)\sum_{j=1}^{N}\|f(x_j(t_i),t_i)\|^2}\Big)^{\frac{1}{2}},\quad \Big(\sum_{i=0}^{T}\frac{\sum_{j=1}^{N}\||\sigma(x_j(t_i),t_i)|-|\hat{\sigma}(x_j(t_i),t_i)|\|^2}{(T+1)\sum_{j=1}^{N}\|\sigma(x_j(t_i),t_i)\|^2}\Big)^{\frac{1}{2}} \tag{44}$$

between the reconstructed $\hat{f},\hat{\sigma}$ in Eq. (2) and the ground-truth f,σ in Eq. (1), respectively. Here, $x_j(t_i)$ is the value of the j^{th} ground-truth trajectory at t_i.

Example 2 Next, we reconstruct a Cox-Ingersoll-Ross (CIR) model, which is a popular finance model that describes the evolution of interest rates:

$$\mathrm{d}X(t)=\big(5-X(t)\big)\mathrm{d}t+\sigma_0\sqrt{X(t)}\mathrm{d}B(t),\ \ t\in[0,2]. \tag{45}$$

Specifically, we are interested in how our learned $\hat{f},\hat{\sigma}$ can approximate the ground-truth $f(X)=5-X$ and $\sigma(X)=\sigma_0\sqrt{X}$ (with σ_0 a constant parameter). Here, we take the timestep $\Delta t=0.05$ in Eq. (29) and the initial condition is $X(0)=2$. For reconstructing f and σ, we compare using our proposed finite-time-point time-decoupled squared W_2 distance Eq. (29) with minimizing a Maximum Mean Discrepancy (MMD) (Briol et al., 2019) and other loss functions given in Appendix E. Our results are shown in Fig. 2. Hyperparameters in the neural networks used for training are the same across all loss functions.

Figure 2a shows the predicted trajectories using our proposed squared W_2 loss function match well with the ground-truth trajectories. Figure 2b, c indicate that, if $\gtrsim 100$ ground-truth trajectories are used, our proposed squared W_2 distance loss yields smaller errors in f,σ as defined in Eq. (44). More specifically, we plot the reconstructed $\hat{f}_\Theta,\hat{\sigma}_\Theta$ by using our squared W_2 loss in Fig. 2d; these reconstructions also match well with the ground-truth values f,σ. When we vary σ_0 in Eq. (45), our proposed finite-time-point time-decoupled W_2 loss function gives the best performance among all loss functions shown in Fig. 2e, f. In Appendix F, instead of using the same initial condition for all trajectories, we sam-

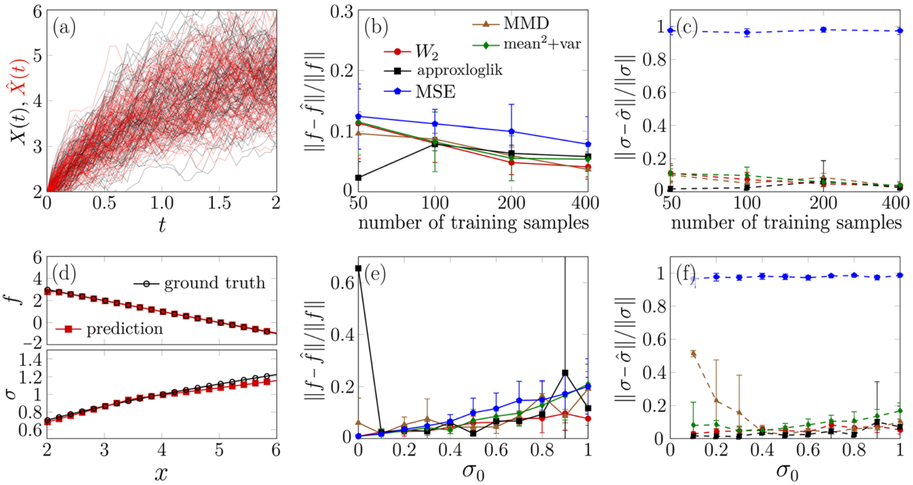

Fig. 2 **a** Ground-truth trajectories and reconstructed trajectories by nSDE using the finite-time-point time-decoupled squared W_2 loss with $\sigma_0=0.5$. **b**, **c** Errors with respect to the numbers of ground-truth trajectories for $\sigma_0=0.5$. **d** Comparison of the reconstructed $\hat{f}_{\Theta_1}(X),\hat{\sigma}_{\Theta_2}(X)$ to the ground-truth functions $f(X),\sigma(X)$ for $\sigma_0=0.5$. **e**, **f** Errors with respect to noise level σ_0 with 200 training samples. Legends for (**c**, **e**, **f**) are the same as the one in (**b**)

ple the initial condition from different distributions and find that the reconstruction errors $f - \hat{f}$ and $\sigma - \hat{\sigma}$ is **not** sensitive to different initial conditions, implying the robustness of using our proposed finite-time-point time-decoupled W_2 loss function with respect to different initial conditions. Also, in Appendix G, we change the number of layers and the number of neurons in each layers for the two neural networks we utilize to parameterize $\hat{f} := \hat{f}(X; \Theta_1), \hat{\sigma} := \hat{\sigma}(X; \Theta_2)$. We find that wider neural networks can lead to smaller errors $f - \hat{f}$ and $\sigma - \hat{\sigma}$.

Next, we reconstruct the Ornstein-Uhlenbeck (OU) process given in Kidger et al. (2021) and in doing so, compare our loss function with the WGAN-SDE method therein and with another recent MMD method.

Example 3 Consider reconstructing the following time-inhomogeneous OU process

$$\mathrm{d}X(t) = \big(0.02t - 0.1X(t)\big)\mathrm{d}t + 0.4\mathrm{d}B(t), \quad t \in [0, 63]. \tag{46}$$

We compare the numerical performance of minimizing Eq. (20) or minimizing Eq. (29) with the WGAN method and using the MMD loss metric. Equation (20) is numerically evaluated using the `ot.emd2` function in the Python Optimal Transport package (Flamary et al., 2021) We take the timestep $\Delta t = 1$ in Eqs. (29) and (20) and the initial condition is taken as $X(0) = 0$. Neural networks with the same number of hidden layers and neurons in each layer are used for all three methods (see Table 1).

In addition to the relative errors in learned $\hat{f}, \hat{\sigma}$, we also compare the runtime and memory usage used by the three methods as a function of the number of ground-truth trajectories used in training.

From Fig. 3a, the distribution of trajectories of the reconstructed SDE found from using our proposed squared W_2 loss Eq. (29) matches well with the distribution of the ground-truth trajectories. Both minimizing Eq. (20) and minimizing Eq. (29) outperform the other two methods in the relative L^2 error of the learned f, σ for all numbers of ground-truth trajectories. Using Eq. (29) as the loss function achieves better accuracy in a shorter computational time than using Eq. (20).

For N_{sample} training samples and N total number of timesteps, the memory cost in using Eq. (29) is $O(N \times N_{\text{sample}})$; however, the number of operations needed is

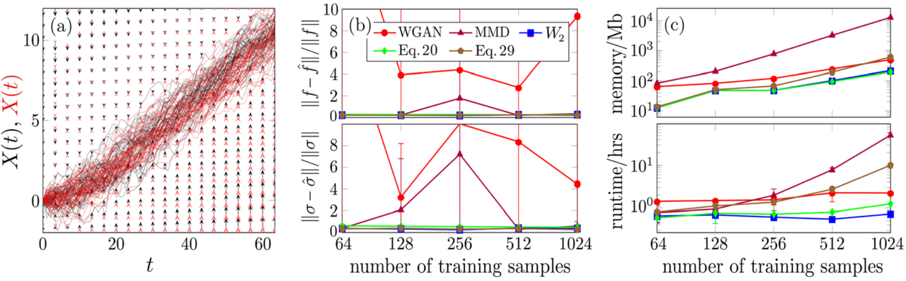

Fig. 3 **a** Ground-truth and reconstructed trajectories using the squared W_2 loss Eq. (29). Black and red curves are ground-truth and reconstructed trajectories, respectively. Black and red arrows indicate $f(x, t)$ and the reconstructed $\hat{f}(x, t)$ at fixed (x, t), respectively. **b** Relative errors in learned $\hat{f}$ and $\hat{\sigma}$, repeated 10 times. Error bars show the standard deviation. **c** Resource consumption with respect to the number of training samples N_{samples}. Memory usage is measured by torch profiler and represents peak memory usage during training. The legend in the (**c**) is the same as the one in (**b**)

$O(N \times N_{\text{sample}} \log N_{\text{sample}})$ because we need to reorder the ground-truth $X(t_i)$ and predicted $\hat{X}(t_i)$ data to obtain the empirical cumulative distributions at every t_i. The memory cost and operations needed in using Eq. (20) are both $O((N \times N_{\text{sample}})^2)$ because a $(N \times N_{\text{sample}}) \times (N \times N_{\text{sample}})$ cost matrix must be evaluated. On the other hand, the MMD method needs to create an $N_{\text{sample}} \times N_{\text{sample}}$ matrix for each timestep and thus the corresponding memory cost and operations needed are at best $O(N \times N_{\text{sample}}^2)$. The WGAN-SDE method needs to create a generator and a discriminator and its training is complex, leading to both a higher memory cost and a larger runtime than our method. When learning SDEs from data, a larger number of ground-truth trajectories leads to higher accuracy (see Appendix H). Overall, our time-decoupled squared W_2 loss, Eq. (29), performs the best in terms of accuracy and efficiency when reconstructing the 1D SDE Eq. (46).

If we consider using stochastic gradient descent (SDG) to minibatch for training, we find that the batch size cannot be set too small, especially when we are using the MMD or Eq. (20) as loss functions, due to the intrinsic noisy nature of trajectories of SDEs. Thus, using our squared W_2 distance loss function given in Eq. (29) can be more efficient overall than using the MMD or Eq. (20) as the loss function. Additional results using the SGD with minibatch for training are given in Appendix H.

Next, we carry out an experiment on reconstructing a 2D correlated geometric Brownian motion. In this 2D reconstruction problem, we will compare the loss functions, Eqs. (20) and (29), the MMD method, and a sliced squared Wasserstein distance method (Kolouri et al., 2018).

Example 4 Consider reconstructing the following 2D correlated geometric Brownian motion that can represent, e.g., values of two correlated stocks (Musiela & Rutkowski, 2006)

$$\begin{aligned} \mathrm{d}X_1(t) =& \mu_1 X_1(t)\mathrm{d}t + \sum_{i=1}^{2} \sigma_{1,i} X_i(t)\mathrm{d}B_i(t), \\ \mathrm{d}X_2(t) =& \mu_2 X_2(t)\mathrm{d}t + \sum_{i=1}^{2} \sigma_{2,i} X_i(t)\mathrm{d}B_i(t) \end{aligned} \tag{47}$$

here, $t \in [0, 2]$, $B_1(t)$ and $B_2(t)$ are independent Brownian processes, $f := (\mu_1 X_1, \mu_2 X_2)$ is a 2D vector, and $\sigma := [\sigma_{1,1}X_1, \sigma_{1,2}X_2; \sigma_{2,1}X_1, \sigma_{2,2}X_2]$ is a 2×2 matrix.

We use $(\mu_1, \mu_2) = (0.1, 0.2)$, $\boldsymbol{\sigma} = [0.2X_1, -0.1X_2; -0.1X_1, 0.1X_2]$, and set the initial condition $(X_1(0), X_2(0)) = (1, 0.5)$. In addition to directly minimizing a 2D decorrelated version of the squared W_2 distance Eq. (29) (denoted as W_2 in Fig. 4c), we consider minimizing a sliced squared W_2 distance as proposed by Kolouri et al. (2018, 2019). Finally, we numerically estimate the W_2 distance Eq. (20) as well as the time-decoupled approximation Eq. (29) using the `ot.emd2` function in the Python Optimal Transport package. Formulae of the above loss functions are given in Appendix E. We keep the neural network hyperparameters the same while minimizing all loss functions. Note that since the SDE has two components, the definition of the relative error in σ is revised to

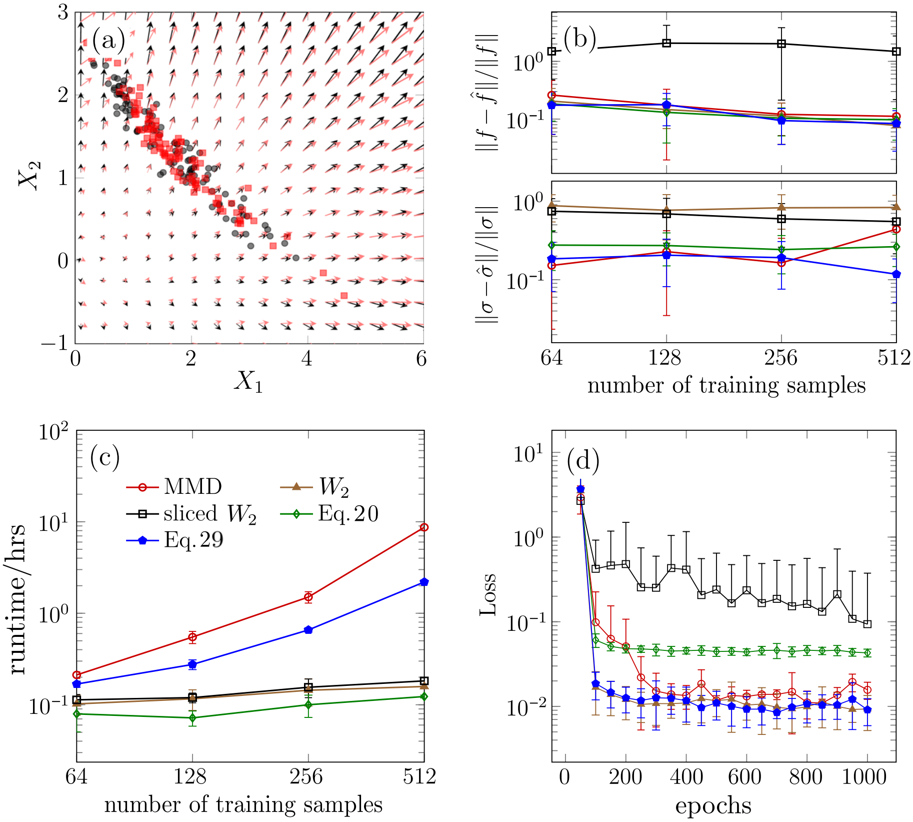

Fig. 4 **a** Black dots and red squares are the ground-truth $(X_1(2), X_2(2))$ and the reconstructed $(\hat{X}_1(2), \hat{X}_2(2))$ found using the rotated squared W_2 loss function, respectively. Black and red arrows indicate, respectively, the vectors $f(X_1, X_2)$ and $\hat{f}(X_1, X_2)$. **b** Relative errors of the learned f and σ. Error bars indicate the standard deviation across ten reconstructions. **c** Runtime of different loss functions with respect to N_{samples}. **d** The decrease of different loss functions with respect to training epochs. The legend for the (**d**) is the same as the one in (**c**)

$$\left[\sum_{i=0}^{T}\frac{\sum_{j=1}^{N}\|\sigma\sigma^T(x_j(t_i),t_i)-\hat{\sigma}\hat{\sigma}^T(x_j(t_i),t_i)\|_F^2}{(T+1)\sum_{j=1}^{N}\|\hat{\sigma}\hat{\sigma}^T(x_j(t_i),t_i)\|_F^2}\right]^{1/2}, \tag{48}$$

where $\|\cdot\|_F$ is the Frobenius norm for matrices.

Figure 4a shows the ground truth and reconstructed coordinates (X_1, X_2) (black dots) and $(\hat{X}_1, \hat{X}_2)$ (red squares) at time $t=2$, along with $f(X_1, X_2)$ (black) and $\hat{f}(X_1, X_2)$ (red). For learning $f$and σ in problem, numerically evaluating Eq. (29) (blue curve) performs better than the MMD method, the loss in Eq. (20), the sliced W_2 distance loss, and the 2D decorrelated squared W_2 loss, as shown in Fig. 4b. Using the sliced W_2 distance yields the poorest performance and least accurate $\hat{f}$ and $\hat{\sigma}$. Using the 2D decorrelated squared W_2 loss function also gives inaccurate $\hat{\boldsymbol{\sigma}}$. Thus, the sliced W_2 distance and the 2D decorrelated squared W_2 loss are not good candidates for learning multivariate SDEs. Numerically estimating Eq. (20) yields poorer performance than numerically estimating Eq. (29) because numerically evaluating the W_2 distance for higher-dimensional empirical distributions is generally less accurate.

From Fig. 4c, we see that the runtime and memory needed to numerically evaluate the time-decoupled Eq. (29) using `ot.emd2` is smaller than those needed for the MMD method, but larger than those needed to numerically estimate Eq. (20). Yet, as shown in Fig. 4d, minimizing Eq. (29) leads to the fastest convergence, potentially requiring fewer epochs when using Eq. (29) as the loss function. An additional comparison of using the two loss functions, the finite-time-point squared W_2 distance Eq. (20) and the finite-time-point time-decoupled squared W_2 distance Eq. (29) is given in Appendix I. Further analysis on how the number of samples and the dimensionality of an SDE dimensionality affects W_2 -based distances in learning multivariate SDEs will be informative.

Finally, to illustrate a biological application of our method, we reconstruct an SDE model developed to describe circadian clocks. Circadian cycles can influence cell gene regulatory dynamics and regulate cell and tissue state dynamics (Gonze, 2011). Intrinsic noise has been hypothesized to play an important role in governing the circadian clock dynamics (Westermark et al., 2009).

Example 5 We formulate an SDE model of circadian cycles derived from adding Brownian noise to an established deterministic model described by five coupled ODEs Goldbeter (1995):

$$
\begin{aligned}
\mathrm{d}M &= \big(v_s\frac{K_I^4}{K_I^4+P_N^4} - v_m\frac{M}{K_m+M}\big)\mathrm{d}t + 0.1M\mathrm{d}B_t^1,\\
\mathrm{d}P_0 &= \big(k_sM - v_1\frac{P_0}{K_1+P_0} + v_2\frac{P_1}{K_2+P_1}\big)\mathrm{d}t + 0.05P_0\mathrm{d}B_t^2,\\
\mathrm{d}P_1 &= \big(v_1\frac{P_0}{K_1+P_0} - v_2\frac{P_1}{K_2+P_1} - v_3\frac{P_1}{K_3+P_1} + v_4\frac{P_2}{K_4+P_2}\big)\mathrm{d}t + 0.1\mathrm{d}B_t^3,\\
\mathrm{d}P_2 &= \big(v_3\frac{P_1}{K_3+P_1} - v_4\frac{P_2}{K_4+P_2} - k_1P_2 + k_2P_N - v_d\frac{P_2}{K_d+P_2}\big)\mathrm{d}t,\\
\mathrm{d}P_N &= \big(k_1P_2 - k_2P_N - v_n\frac{P_N}{K_n+P_N}\big)\mathrm{d}t + 0.01\mathrm{d}B_t^4,\quad t\in[0,50].
\end{aligned}
\tag{49}
$$

In Eq. (49), *M* describes the concentration of mRNA, P_0 is the concentration of native protein (*per*), P_1, P_2 represent concentrations of two different forms of phosphorylated protein *per* with one or two phosphorylated sites, and P_N quantifies the concentration of nuclear *per*. The parameters $K_1 = 2\mu\text{mol}, K_2 = 2\mu\text{mol}, K_3 = 2\mu\text{mol}, K_4 = 2\mu\text{mol}, K_n, K_I = 1\mu\text{mol}$, and $K_m = 0.5\mu\text{mol}$ are the corresponding Michaelis-Menten constants. Reaction rates are represented by $v_s = 0.76\mu\text{mol/hr}, v_1 = 3.2\mu\text{mol/hr}, v_2 = 1.58\mu\text{mol/hr}, v_3 = 5\mu\text{mol/hr}, v_4 = 2.5\mu\text{mol/hr}$, $v_m = 0.65\mu\text{mol/hr}, v_d = 0.95\mu\text{mol/hr}$, and $k_s = 0.38/\text{hr}, k_1 = 1.9/\text{hr}, k_2 = 1.3/\text{hr}$. The dynamics involve four independent Brownian noises described as $\mathrm{d}B_t^i, i = 1, 2, 3, 4$.

We plot the ground truth trajectories and the trajectories generated by neural SDEs trained through minimizing the time-decoupled W_2-distance loss function (in Appendix E) in Fig. 5. The training details are given in Table 1. For simplicity, we plot the mRNA concentration *M*, the naive protein concentration P_0, and the nuclear protein concentration P_N, which all display periodic fluctuations over time. The reconstructed trajectories by our Wasserstein-distance SDE approach can accurately reproduce the intrinsically noisy peri-

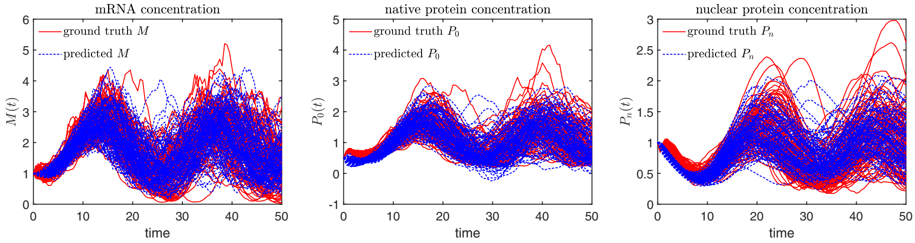

Fig. 5 The reconstructed trajectories using our Wasserstein-distance SDE reconstruction approach compared to the ground truth trajectories obtained by simulating Eqs. (49). For simplicity, we plot the ground truth and reconstructed trajectories of the concentrations of mRNA, native protein, and nuclear protein. The initial condition is set as $(M(0), P_0(0), P_1(0), P_2(0), P_N(0)) = (1, 0.5, 2, 0, 1)$ (units: μmol) for all trajectories

odic changes in the mRNA and protein concentrations. More analysis of the applicability of minimizing the time-decoupled W_2-distance loss function to train higher-dimensional neural SDEs requires further investigation.

6 Summary and conclusions

In this paper, we analyzed the squared W_2 distance between two probability distributions associated with two SDEs and proposed a novel method for efficiently learning SDEs from data by minimizing squared W_2 distances as loss functions. Upon performing numerical experiments, we found that our proposed finite-time-point time-decoupled squared W_2 distance loss function, Eq. (29), is superior than many other recently developed machine-learning and statistical approaches to SDE reconstruction.

A number of extensions are apparent. First, one can further investigate applying the squared W_2 loss to the reconstruction of high-dimensional SDEs. Whether the Wasserstein distance can serve as upper bounds for the errors $f - \hat{f}$ and $\sigma - \hat{\sigma}$ is also an intriguing question as its resolution will determine whether minimizing the squared Wasserstein distance is sufficient for learning SDEs. Another promising area worthy of study is the extension of the squared W_2 distance loss function to the reconstruction of general Lévy processes that include jumps in the trajectories. Finally, how to take into account extrinsic noise, e.g., errors in the measurement of time-series data, could be a prospective research field.

Appendix A: Proof to theorem 1

Here, we shall provide a proof to Theorem 1. First, note that $\tilde{X}(t)$ defined in Eq. (10) has the same distribution as that of $\hat{X}(t)$ defined in Eq. (2). Therefore, by definition, if we let π in Eq. (4) to be the joint distribution of $(X, \tilde{X})$, then

$$W_2(\mu, \hat{\mu}) \leq \left(\mathbb{E}\left[\int_0^T |\tilde{X}(t) - X(t)|^2 \mathrm{d}t\right]\right)^{1/2}. \tag{A1}$$

Next, we provide a bound for $\mathbb{E}\big[\int_0^T |\tilde{X}(t) - X(t)|^2 \mathrm{d}t\big]^{1/2}$ by the mean value theorem for *f* and *g*. Note that the standard Brownian motion *B*(*t*) in Eq. (10) is identical to that in Eq. (1) and

$$\begin{aligned}\mathrm{d}\big(X(t) - \tilde{X}(t)\big) =& \partial_x f\big(\eta_1(X(t), \tilde{X}(t), t), t\big) \cdot (X(t) - \tilde{X}(t))\mathrm{d}t \\ &+ \partial_x \sigma\big(\eta_2(X(t), \tilde{X}(t)), t\big) \cdot (X(t) - \tilde{X}(t))\mathrm{d}B(t) \\ &+ (f - \hat{f})(\tilde{X}(t), t)\mathrm{d}t + (\sigma - \hat{\sigma})(\tilde{X}(t), t))\mathrm{d}B(t).\end{aligned} \tag{A2}$$

where $\eta_1(x_1, x_2), \eta_2(x_1, x_2)$ are defined in Eq. (11) such that their values are in (x_1, x_2).

Applying Itô's formula to $[X(t) - \tilde{X}(t)]/H(0;t)$, where H(0, t) is defined in Eq. (12) we find

$$\begin{aligned}\mathrm{d}\left(\frac{X(t) - \tilde{X}(t)}{H(0;t)}\right) =& \frac{1}{H(0;t)}\Big[(f - \hat{f})(\tilde{X}(t), t)\mathrm{d}t + \partial_x \sigma\big(\eta_2(X, \tilde{X}), t\big) \cdot (\sigma - \hat{\sigma})(\tilde{X}(t), t)\mathrm{d}t\Big] \\ &+ \frac{1}{H(0;t)}\Big[(\sigma - \hat{\sigma})(\tilde{X}(t), t)\mathrm{d}B(t)\Big].\end{aligned} \tag{A3}$$

Integrating both sides from 0 to *t*, we obtain

$$\begin{aligned}X(t) - \tilde{X}(t) =& \int_0^t H(s;t)\Big[(f - \hat{f})(\tilde{X}(s), s) + \partial_x \sigma\big(\eta_2(X, \tilde{X}), s\big) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\Big]\mathrm{d}s \\ &+ \int_0^t H(s;t) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\mathrm{d}B(s).\end{aligned} \tag{A4}$$

By invoking Itô isometry and observing that $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$, we deduce

$$\begin{aligned}\mathbb{E}\big[\big(X(t) - \tilde{X}(t)\big)^2\big] \le& 3\mathbb{E}\Big[\Big(\int_0^t H(s;t) \cdot (f - \hat{f})(\tilde{X}(s), s)\mathrm{d}s\Big)^2\Big] \\ &+ 3\mathbb{E}\Big[\Big(\int_0^t H(s;t) \cdot (\partial_x \sigma\big(\eta_2(X, \tilde{X}), s\big) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\mathrm{d}s\Big)^2\Big] \\ &+ 3\mathbb{E}\Big[\Big(\int_0^t H(s;t) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\mathrm{d}B(s)\Big)^2\Big] \\ \le& 3\mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big] \times \mathbb{E}\Big[\int_0^T (f - \hat{f})^2(\tilde{X}(s), s)\mathrm{d}s\Big] \\ &+ 3\mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big] \times \mathbb{E}\Big[\int_0^t \Big(\partial_x \sigma\big(\eta_2(X, \tilde{X}), s\big) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\Big)^2 \mathrm{d}s\Big] \\ &+ 3\mathbb{E}\Big[\int_0^T H^2(s;t) \cdot (\sigma - \tilde{\sigma})^2(\tilde{X}(s), s)\mathrm{d}s\Big] \\ \le& 3\mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big] \times \mathbb{E}\Big[\int_0^t (f - \hat{f})^2(\tilde{X}(s), s)\mathrm{d}s\Big] \\ &+ 3\mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big] \times \mathbb{E}\Big[\int_0^t \Big(\partial_x \sigma\big(\eta_2(X, \tilde{X}), s\big) \cdot (\sigma - \hat{\sigma})(\tilde{X}(s), s)\Big)^2 \mathrm{d}s\Big] \\ &+ 3\left(\mathbb{E}\Big[\int_0^t H^4(s;t)\mathrm{d}s\Big]\right)^{1/2} \times \left(\mathbb{E}\Big[\int_0^t (\sigma - \hat{\sigma})^4(\tilde{X}(s), s)\mathrm{d}s\Big]\right)^{1/2}.\end{aligned} \tag{A5}$$

Finally, we conclude that

$$
\begin{aligned}
W_2^2(\mu,\tilde{\mu}) \le & \int_0^T \mathbb{E}\big[\big(X(t)-\tilde{X}(t)\big)^2\big]\mathrm{d}t\\
\le & 3\int_0^T \mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big]\mathrm{d}t \times \mathbb{E}\Big[\int_0^T (f-\hat{f})^2(\tilde{X}(s),s)\mathrm{d}s)\Big]\\
& + 3\int_0^T \mathbb{E}\Big[\int_0^t H^2(s;t)\mathrm{d}s\Big]\mathrm{d}t \times \mathbb{E}\Big[\int_0^T (\partial_x\sigma(\eta_2(X,\tilde{X}),s)\cdot(\sigma-\hat{\sigma})(\tilde{X}(s),s))^2\mathrm{d}s)\Big]\\
& + 3\int_0^T \bigg(\mathbb{E}\Big[\int_0^t H^4(s;t)\mathrm{d}s\Big]\bigg)^{1/2}\mathrm{d}t \times \bigg(\mathbb{E}\Big[\int_0^T (\sigma-\hat{\sigma})^4(\tilde{X}(s),s)\mathrm{d}s)\Big]\bigg)^{1/2},
\end{aligned}
\tag{A6}
$$

which proves Theorem 1.

Appendix B: Single-trajectory MSE and KL divergence

We shall first show that using the single-trajectory MSE tends to fit the mean process $\mathbb{E}[X(t)]$ and make noise diminish, which indicates that the MSE is not a good loss function when one wishes to fit σ in Eq. (1).

For two *independent* d-dimensional stochastic processes $\{\boldsymbol{X}(t)\}_{t=0}^T, \{\hat{\boldsymbol{X}}(t)\}_{t=0}^T$ as solutions to Eqs. (1) and (2) with appropriate $f, \hat{f}$ and $\sigma, \hat{\sigma}$, let $\mathbb{E}[\boldsymbol{X}]$ represent the trajectory of mean values of $\boldsymbol{X}(t)$, i.e., $\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}(t)]$. We have

$$
\mathbb{E}\big[\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2\big] = \mathbb{E}\big[\,\|\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}]\|^2\,\big] + \mathbb{E}\big[\|\hat{\boldsymbol{X}}-\mathbb{E}[\boldsymbol{X}]\|^2\big] - 2\mathbb{E}\Big[\int_0^T \big(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}], \hat{\boldsymbol{X}}-\mathbb{E}[\boldsymbol{X}]\big)\,\mathrm{d}t\Big], \tag{A7}
$$

where$\|\boldsymbol{X}\|^2 := \int_0^T |\boldsymbol{X}|_2^2\,\mathrm{d}t, |\cdot|_2$denotesthe$\ell^2$normofavector,and$(\cdot,\cdot)$istheinnerproduct oftwo$d$-dimensionalvectors.Inviewoftheindependencebetween$\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}]and\hat{\boldsymbol{X}}-\mathbb{E}[\boldsymbol{X}]$, we have $\mathbb{E}\big[\big(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}], \hat{\boldsymbol{X}}-\mathbb{E}[\boldsymbol{X}]\big)\big] = \mathbb{E}\big[\,(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])\,\big]\cdot\mathbb{E}\big[(\hat{\boldsymbol{X}}-\mathbb{E}[\boldsymbol{X}])\big] = 0$, and

$$
\mathbb{E}\big[\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2\big] \ge \mathbb{E}\big[\,\|\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}]\|^2\,\big]. \tag{A8}
$$

Therefore, the optimal $\hat{\boldsymbol{X}}$ that minimizes the MSE is $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}]$, which indicates that the MSE tends to fit the mean process $\mathbb{E}[\boldsymbol{X}]$ and make noise diminish. This is not desirable when one wishes to fit a nonzero σ in Eq. (1).

The KL divergence, in some cases, will diverge and thus is not suitable for being used as a loss function. Here, we provide a simple intuitive example when the KL divergence fail. If we consider the degenerate case when $\mathrm{d}X(t) = \mathrm{d}t, \mathrm{d}\hat{X}(t) = (1-\epsilon)\mathrm{d}t, t \in [0,T]$, then $D_{KL}(\mu,\hat{\mu}) = \infty$ no matter how small $\epsilon \ne 0$ is because $\mu, \hat{\mu}$ has different and degenerate support. However, from Theorem 1, $\lim\limits_{\epsilon\to 0} W_2(\mu,\hat{\mu}) = 0$. Therefore, the KL divergence cannot effectively measure the similarity between $\mu, \hat{\mu}$. Overall, the squared W_2 distance is a better metric than some of the commonly used loss metrics such as the MSE or the KL divergence.

Appendix C: Proof to theorem 2

Here, we shall prove Theorem 2. We denote

$$\Omega_N := \{\boldsymbol{Y}(t)|\boldsymbol{Y}(t) = \boldsymbol{Y}(t_i),\ t \in [t_i, t_{i+1}),\ i < N-1;\ \boldsymbol{Y}(t) = \boldsymbol{Y}(t_i),\ t \in [t_i, t_{i+1}],\ i = N-1\} \tag{A9}$$

to be the space of piecewise functions. We also define the space

$$\tilde{\Omega}_N := \{\boldsymbol{Y}_1(t) + \boldsymbol{Y}_2(t), \boldsymbol{Y}_1 \in C([0,T];\mathbb{R}^d), \boldsymbol{Y}_2 \in \Omega_N\}. \tag{A10}$$

$\tilde{\Omega}_N$ is also a separable metric space because both $\big(C([0,T];\mathbb{R}^d), \|\cdot\|\big)$ and $\big(\Omega_N, \|\cdot\|\big)$ are separable metric spaces. Furthermore, both the embedding mapping from $C([0,T];\mathbb{R}^d)$ to $\tilde{\Omega}_N$ and the embedding mapping from Ω_N to $\tilde{\Omega}_N$ preserves the $\|\cdot\|$ norm. Then, the two embedding mappings are measurable, which enables us to define the measures on $\mathcal{B}(\tilde{\Omega}_N)$ induced by the measures $\mu, \hat{\mu}$ on $\mathcal{B}\big(C([0,T];\mathbb{R}^d)\big)$ and the measures $\mu_N, \hat{\mu}_N$ on $\mathcal{B}(\Omega_N)$. For notational simplicity, we shall still denote those induced measures by $\mu, \hat{\mu}, \mu_N, \hat{\mu}_N$.

Therefore, the inequality Eq. (17) is a direct result of the triangular inequality for the Wasserstein distance (Clement & Desch, 2008) because $\boldsymbol{X}, \boldsymbol{X}_N, \hat{\boldsymbol{X}}, \hat{\boldsymbol{X}}_N \in \tilde{\Omega}_N$.

Next, we shall prove Eq. (19) when $\boldsymbol{X}(t), \hat{\boldsymbol{X}}(t)$ are solutions to SDEs Eqs. (1) and (2). Because $\boldsymbol{X}_N(t)$ is the projection to $\boldsymbol{X}(t)$, the squared $W_2^2(\mu, \mu_N)$ can be bounded by

$$W_2^2(\mu, \mu_N) \leq \sum_{i=1}^{N}\int_{t_{i-1}}^{t_i} \mathbb{E}\big[\big|\boldsymbol{X}(t) - \boldsymbol{X}_N(t)\big|_2^2\big]\mathrm{d}t = \sum_{i=1}^{N}\int_{t_{i-1}}^{t_i} \sum_{\ell=1}^{d}\mathbb{E}\big[\big(X_\ell(t) - X_{N,\ell}(t)\big)^2\big]\mathrm{d}t \tag{A11}$$

For the first inequality above, we choose a specific *coupling*, i.e. the coupled distribution, π of μ, μ_N that is essentially the "original" probability distribution. To be more specific, for an abstract probability space $(\Omega, \mathcal{A}, p)$ associated with $\boldsymbol{X}$, μ and μ_N can be characterized by the *pushforward* of p via $\boldsymbol{X}$ and $\boldsymbol{X}_N$ respectively, i.e., $\mu = \boldsymbol{X}_* p$, defined by $\forall A \in \mathcal{B}\big(\tilde{\Omega}_N\big)$, elements in the Borel σ-algebra of $\tilde{\Omega}_N$,

$$\mu(A) = \boldsymbol{X}_* p(A) := p\big(\boldsymbol{X}^{-1}(A)\big), \tag{A12}$$

where $\boldsymbol{X}$ is interpreted as a measurable map from Ω to $\tilde{\Omega}_N$, and $\boldsymbol{X}^{-1}(A)$ is the preimage of A under $\boldsymbol{X}$. Then, the coupling π is defined by

$$\pi = (\boldsymbol{X}, \boldsymbol{X}_N)_* p, \tag{A13}$$

where $(\boldsymbol{X}, \boldsymbol{X}_N)$ is interpreted as a measurable map from Ω to $\tilde{\Omega}_N \times \tilde{\Omega}_N$. One can readily verify that the marginal distributions of π are μ and μ_N respectively. Recall that s represents the dimension of the standard Brownian motions in the SDEs.

For each $\ell = 1, ..., d$, we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\int_{t_{i-1}}^{t_i}\mathbb{E}\big[\big(X_\ell(t)-X_{N,\ell}(t)\big)^2\big]\mathrm{d}t\\
&\quad\le (s+1)\left[\sum_{i=1}^{N}\int_{t_{i-1}}^{t_i}\left(\mathbb{E}\Big[\Big(\int_{t_i}^{t} f_\ell(\boldsymbol{X}(r),r)\mathrm{d}r\Big)^2\Big]+\mathbb{E}\Big[\Big(\int_{t_i}^{t}\sum_{j=1}^{s}\sigma_{\ell,j}(\boldsymbol{X}(r),r)\mathrm{d}B_j(r)\Big)^2\Big]\right)\mathrm{d}t\right]\\
&\quad\le (s+1)\sum_{i=1}^{N}\left((\Delta t)^2\mathbb{E}\Big[\int_{t_{i-1}}^{t_i} f_\ell^2\mathrm{d}t\Big]+\Delta t\sum_{j}\mathbb{E}\Big[\int_{t_{i-1}}^{t_i}\sigma_{\ell,j}^2\mathrm{d}t\Big]\right)
\end{aligned}
\tag{A14}
$$

The first inequality follows from the observation that $\big(\sum_{i=1}^n a_i\big)^2 \le n(\sum_{i=1}^n a_i^2)$ and application of this observation to the integral representation of $\boldsymbol{X}(t)$. Summing over ℓ, we have

$$
\Big(\sum_{i=1}^{N}\int_{t_{i-1}}^{t_i}\mathbb{E}\big[\big|\boldsymbol{X}(t)-\boldsymbol{X}_N(t)\big|_2^2\big]\mathrm{d}t\Big)^{1/2}\le\sqrt{s+1}\,\Big(F(\Delta t)^2+\Sigma\Delta t\Big)^{1/2}
\tag{A15}
$$

Similarly, $W_2(\hat{\mu},\hat{\mu}_N)$ can be bounded by

$$
W_2(\hat{\mu},\hat{\mu}_N)\le\sqrt{s+1}\sqrt{\hat{F}(\Delta t)^2+\hat{\Sigma}\Delta t}.
\tag{A16}
$$

Substituting Eqs. (A15) and (A16) into Eq. (17), we have proved Eq. (19). This completes the proof of Theorem 2.

Appendix D: Proof to Theorem 3

We now give a proof to Theorem 3. First, we notice that

$$
\mathbb{E}\left[\big|\boldsymbol{X}(t)-\hat{\boldsymbol{X}}(t)\big|_2^2\right]\le 2\big(FT+\hat{F}T+\Sigma+\hat{\Sigma}\big)<\infty,\ \ \forall t\in[0,T]
\tag{A17}
$$

where $F,\hat{F},\Sigma,\hat{\Sigma}$ are defined in Eq. (18). We denote

$$
M:=\max_{t\in[0,T]}W_2\big(\mu(t),\hat{\mu}(t)\big)\le 2\big(FT+\hat{F}T+\Sigma+\hat{\Sigma}\big).
\tag{A18}
$$

By applying Theorem 2 with $N=1$, the bound

$$
\begin{aligned}
&\inf_{\pi_i}\sqrt{\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t}-\sqrt{(s+1)\Delta t}\left(\sqrt{F_i\Delta t+\Sigma_i}+\sqrt{\hat{F}_i\Delta t+\hat{\Sigma}_i}\right)\\
&\quad\le W_2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i)\\
&\quad\le\inf_{\pi_i}\sqrt{\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t}+\sqrt{(s+1)\Delta t}\left(\sqrt{F_i\Delta t+\Sigma_i}+\sqrt{\hat{F}_i\Delta t+\hat{\Sigma}_i}\right).
\end{aligned}
\tag{A19}
$$

holds true for all $i = 1, 2, ..., N-1$. In Eq. (A19),

$$F_i := \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d} f_\ell^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big] < \infty,\ \Sigma_i := \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d}\sum_{j=1}^{s} \sigma_{\ell,j}^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big] < \infty,$$
$$\hat{F}_i := \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d} \hat{f}_\ell^2(\hat{\boldsymbol{X}}(t), t)\mathrm{d}t\Big] < \infty,\ \hat{\Sigma}_i := \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d}\sum_{j=1}^{s} \hat{\sigma}_{\ell,j}^2(\hat{\boldsymbol{X}}(t), t)\mathrm{d}t\Big] < \infty, \tag{A20}$$

which results from

$$\sum_{i=0}^{N-1} F_i = F < \infty,\ \sum_{i=0}^{N-1} \hat{F}_i = \hat{F} < \infty,\ \sum_{i=0}^{N-1} \Sigma_i = \Sigma < \infty,\ \sum_{i=0}^{N-1} \hat{\Sigma}_i = \hat{\Sigma} < \infty, \tag{A21}$$

where $F, \hat{F}, \Sigma, \hat{\Sigma}$ are defined in Eq. (18). Squaring the inequality (A19), we have

$$\begin{aligned}
W_2^2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_i) \leq & \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t \\
& + 2\inf_{\pi_i} \sqrt{\mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]}\sqrt{(s+1)\Delta t}\left(\sqrt{F_i\Delta t + \Sigma_i} + \sqrt{\hat{F}\Delta t + \hat{\Sigma}_i}\right) \\
& + 2(s+1)\Delta t\big(F_i\Delta t + \Sigma_i + \hat{F}_i\Delta t + \hat{\Sigma}_i\big), \\
W_2^2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_i) \geq & \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i) - \hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t \\
& - 2W_2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_i)\sqrt{(s+1)\Delta t}\left(\sqrt{F_i\Delta t + \Sigma_i} + \sqrt{\hat{F}\Delta t + \hat{\Sigma}_i}\right) \\
& - 2(s+1)\Delta t\big(F_i\Delta t + \Sigma_i + \hat{F}_i\Delta t + \hat{\Sigma}_i\big)
\end{aligned} \tag{A22}$$

Specifically, from Eqs. (A18) and (A19),

$$W_2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_i) \leq \sqrt{\Delta t}\left[M + \sqrt{s+1}\Big(\sqrt{FT + \Sigma} + \sqrt{\hat{F}T + \hat{\Sigma}}\Big)\right] := \tilde{M}\sqrt{\Delta t}, \quad \tilde{M} < \infty \tag{A23}$$

Summing over $i = 1, ..., N-1$ for both inequalities in Eq. (A22) and noting that $\Delta t = \frac{T}{N}$, we conclude

$$\begin{aligned}\sum_{i=1}^{N-1} W_2^2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i) \leq & \sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\,\Delta t \\ & + 2M\Delta t\sqrt{s+1}\sum_{i=1}^{N-1}\left(\sqrt{F_i\Delta t+\Sigma_i}+\sqrt{\hat{F}_i\Delta t+\hat{\Sigma}_i}\right) \\ & + 2\Delta t(s+1)\big(F\Delta t+\Sigma+\hat{F}\Delta t+\hat{\Sigma}\big), \qquad \text{(A24)} \\ \leq & \sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\Big]\,\Delta t \\ & + 2(s+1)\Delta t\big(F\Delta t+\Sigma+\hat{F}\Delta t+\hat{\Sigma}\big) \\ & + M\sqrt{(s+1)\Delta t}\,\big((F+\hat{F}+2T)\sqrt{\Delta t}+\Sigma+\hat{\Sigma}+2T\big)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{N-1} W_2^2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i) \geq & \sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)|_2^2\big]\Delta t \\ & - 2\tilde{M}\Delta t\sqrt{s+1}\sum_{i=1}^{N-1}\left(\sqrt{F_i\Delta t+\Sigma_i}+\sqrt{\hat{F}_i\Delta t+\hat{\Sigma}_i}\right) \\ & - 2(s+1)\Delta t\,\big(F\Delta t+\Sigma+\hat{F}\Delta t+\hat{\Sigma}\big), \qquad \text{(A25)} \\ \geq & \sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)|_2^2\big]\Delta t \\ & - 2(s+1)\Delta t\,\big(F\Delta t+\Sigma+\hat{F}\Delta t+\hat{\Sigma}\big) \\ & - \tilde{M}\sqrt{(s+1)\Delta t}\,\big((F+\hat{F})\Delta t+\Sigma+\hat{\Sigma}+2T\big).\end{aligned}$$

Eqs. (A24) and (A25) indicate that as $N\to\infty$,

$$\sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[\big|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)\big|_2^2\big]\Delta t - \sum_{i=1}^{N-1} W_2^2(\boldsymbol{\mu}_i,\hat{\boldsymbol{\mu}}_i)\to 0, \qquad \text{(A26)}$$

Suppose $0=t_0^1<t_1^1<...<t_{N_1}^1=T; 0=t_0^2<t_1^2<...<t_{N_2}^2=T$ to be two sets of grids on [0, *T*]. We define a third set of grids $0=t_0^3<...<t_{N_3}^3=T$ such that $\{t_0^1,...,t_{N_1}^1\}\cup\{t_0^2,...,t_{N_2}^2\}=\{t_0^3,...,t_{N_3}^3\}$. Let $\delta t:=\max\{\max_i(t_{i+1}^1-t_i^1),\max_j(t_{j+1}^2-t_j^2),\max_k(t_{k+1}^3-t_k^3)\}$. We denote $\mu(t_i^1)$ and $\hat{\mu}(t_i^1)$ to be the probability distribution of $\boldsymbol{X}(t_i^s)$ and $\hat{\boldsymbol{X}}(t_i^s)$, $s=1,2,3$, respectively. We will prove

$$\Big|\sum_{i=0}^{N_1-1} W_2^2\big(\mu(t_i^1),\hat{\mu}(t_i^1)\big)(t_{i+1}^1-t_i^1) - \sum_{i=0}^{N_3-1} W_2^2\big(\mu(t_i^3),\hat{\mu}(t_i^3)\big)(t_{i+1}^3-t_i^3)\Big|\to 0, \qquad \text{(A27)}$$

as $\delta t \to 0$.

First, suppose in the interval (t_i^1, t_{i+1}^1), we have $t_i^1 = t_\ell^3 < t_{\ell+1} < ... < t_{\ell+s}^3 = t_{i+1}^1, s \geq 1$, then for $s > 1$, since $t_{i+1}^1 - t_i^1 = \sum_{k=\ell}^{\ell+s-1}(t_{k+1}^3 - t_k^3)$, we have

$$\begin{aligned}
&\Big|W_2^2\big(\mu(t_i^1), \hat{\mu}(t_i^1)\big)(t_{i+1}^1 - t_i^1) - \sum_{k=\ell}^{\ell+s-1} W_2^2\big(\mu(t_k^3)), \hat{\mu}(t_i^3)\big)(t_{k+1}^3 - t_k^3)\Big| \\
&\quad \leq \sum_{k=\ell+1}^{\ell+s-1} \Big|W_2\big(\mu(t_i^1), \hat{\mu}(t_i^1)\big) - W_2\big(\hat{\mu}(t_i^3), \hat{\mu}(t_k^3)\big)\Big| \\
&\qquad \times \Big(W_2\big(\mu(t_i^1), \hat{\mu}(t_i^1)\big) + W_2\big(\mu(t_k^3), \hat{\mu}(t_k^3)\big)\Big)(t_{k+1}^3 - t_k^3).
\end{aligned} \tag{A28}$$

On the other hand, because we can take a specific coupling π^* to be the joint distribution of $(\boldsymbol{X}(t_i^1), \boldsymbol{X}(t_k^3))$,

$$\begin{aligned}
W_2(\mu(t_i^1), \mu(t_k^3)) &\leq \Big(\mathbb{E}\big[|\boldsymbol{X}(t_k^3) - \boldsymbol{X}(t_i^1)|_2^2\big]\Big)^{1/2} \\
&\leq \sqrt{s+1}\, \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d} f_\ell^2(\boldsymbol{X}(t), t)\mathrm{d}t + \sum_{\ell=1}^{d}\sum_{j=1}^{s} \sigma_{\ell,j}^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big]^{1/2}
\end{aligned} \tag{A29}$$

Similarly, we have

$$W_2\big(\hat{\mu}(t_i^1), \hat{\mu}(t_k^3)\big) \leq \sqrt{s+1}\, \mathbb{E}\Big[\int_{t_i}^{t_{i+1}} \sum_{\ell=1}^{d} \hat{f}_\ell^2(\boldsymbol{X}(t), t)\mathrm{d}t + \sum_{\ell=1}^{d}\sum_{j=1}^{s} \hat{\sigma}_{\ell,j}^2(\boldsymbol{X}(t), t)\mathrm{d}t\Big]^{1/2} \tag{A30}$$

From the triangular inequality of the Wasserstein distance, we find

$$\Big|W_2\big(\mu(t_i^1), \hat{\mu}(t_i^1)\big) - W_2\big(\mu(t_k^3), \hat{\mu}(t_k^3)\big)\Big| \leq W_2\big(\mu(t_i^1), \mu(t_k^3)\big) + W_2\big(\hat{\mu}(t_i^1), \hat{\mu}(t_k^3)\big). \tag{A31}$$

Substituting Eq. (A31) into Eq. (A28), we conclude that

$$\begin{aligned}
&\Big|W_2^2\big(\mu(t_i^1), \hat{\mu}(t_i^1)\big)(t_{i+1}^1 - t_i^1) - \sum_{k=\ell}^{\ell+s-1} W_2^2\big((\mu(t_k^3), \hat{\mu}(t_k^3)\big)(t_{k+1}^3 - t_k^3)\Big| \\
&\qquad \leq 2M(t_{i+1}^1 - t_i^1)\left(\sqrt{F_i\delta t + \Sigma_i} + \sqrt{\hat{F}_i\delta t + \hat{\Sigma}_i}\right).
\end{aligned} \tag{A32}$$

When the conditions in Eq. (33) hold true, we use Eq. (A32) in Eq. (A27) to find

$$\begin{aligned}&\Big|\sum_{i=0}^{N_1-1} W_2^2\big(\mu(t_i^1),\hat{\mu}(t_i^1)\big)(t_{i+1}^1-t_i^1)-\sum_{i=0}^{N_3-1} W_2^2\big(\mu(t_i^3),\hat{\mu}(t_i^3)\big)(t_{i+1}^3-t_i^3)\Big|\\&\qquad\leq 2MT\max_i\left(\sqrt{F_i\delta t+\Sigma_i}+\sqrt{\hat{F}_i\delta t+\hat{\Sigma}_i}\right)\to 0\end{aligned}\tag{A33}$$

as $\delta t\to 0$. Similarly,

$$\begin{aligned}&\Big|\sum_{i=0}^{N_2-1} W_2^2\big(\mu(t_i^2),\hat{\mu}(t_i^2)\big)(t_{i+1}^2-t_i^2)-\sum_{i=0}^{N_3-1} W_2^2\big(\mu(t_i^3),\hat{\mu}(t_i^3)\big)(t_{i+1}^3-t_i^3)\Big|\\&\qquad\leq 2MT\max_i\left(\sqrt{F_i\delta t+\Sigma_i}+\sqrt{\hat{F}_i\delta t+\hat{\Sigma}_i}\right)\to 0\end{aligned}\tag{A34}$$

as $\delta t\to 0$. Thus,

$$\Big|\sum_{i=0}^{N_1-1} W_2^2\big(\mu(t_i^1),\hat{\mu}(t_i^1)\big)(t_{i+1}^1-t_i^1)-\sum_{i=0}^{N_2-1} W_2^2\big(\mu(t_i^2),\hat{\mu}(t_i^2)\big)(t_{i+1}^2-t_i^2)\Big|\to 0\tag{A35}$$

as $\delta t\to 0$, which implies the limit

$$\lim_{N\to\infty}\sum_{i=1}^{N-1}\inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\big|\boldsymbol{X}(t_i^1)-\hat{\boldsymbol{X}}(t_i^1)\big|_2^2\Big]\,(t_i^1-t_{i-1}^1)=\lim_{N\to\infty}\sum_{i=1}^{N-1}W_2^2\big(\mu(t_i^1),\hat{\mu}(t_i^1)\big)(t_i^1-t_{i-1}^1)\tag{A36}$$

exists. This completes the proof of Theorem 3.

Appendix E: Definition of different loss metrics used in the examples

Six loss functions for 1D cases were considered:

1. The squared Wasserstein-2 distance (Eq. (20))

$$W_2^2(\mu_N^{\mathrm{e}},\hat{\mu}_N^{\mathrm{e}}),$$

where μ_N^{e} and $\hat{\mu}_N^{\mathrm{e}}$ are the empirical distributions of the vector $(X(t_1),...X(t_{N-1}))$ and $(\hat{X}(t_1),...,\hat{X}(t_{N-1}))$, respectively. It is estimated by

$$W_2^2(\mu_N^{\mathrm{e}},\hat{\mu}_N^{\mathrm{e}})\approx \mathrm{ot.emd2}\Big(\frac{1}{M}\boldsymbol{I}_M,\frac{1}{M}\boldsymbol{I}_M,\boldsymbol{C}\Big),\tag{A37}$$

where $\mathrm{ot.emd2}$ is the function for solving the earth movers distance problem in the ot package of Python, M is the number of ground-truth and predicted trajectories, $\boldsymbol{I}_\ell$ is an M-dimensional vector whose elements are all 1, and $\boldsymbol{C}\in\mathbb{R}^{M\times M}$ is a matrix with entries

$(\boldsymbol{C})_{ij} = (X_N^i - \hat{X}_N^j)_2^2$. X_N^i is the vector of the values of the i^{th} ground-truth trajectory at time points $t_1, ..., t_{N-1}$, and $\hat{X}_N^j$ is the vector of the values of the j^{th} predicted trajectory at time points $t_1, ..., t_{N-1}$.

2. The squared time-decoupled Wasserstein-2 distance averaged over each time step (Eq. (29)):

$$\tilde{W}_2^2(\mu_N, \hat{\mu_N}) = \sum_{i=1}^{N-1} W_2^2(\mu_N^{\text{e}}(t_i), \hat{\mu}_N^{\text{e}}(t_i))\Delta t$$

, where Δt is the time step and W_2 is the Wasserstein-2 distance between two empirical distributions $\mu_N^{\text{e}}(t_i), \hat{\mu}_N^{\text{e}}(t_i)$. These distributions are calculated by the samples of the trajectories of $X(t), \hat{X}(t)$ at a given time step $t = t_i$, respectively.

3. Mean squared error (MSE) between the trajectories, where M is the total number of the ground-truth and prediction trajectories. $X_{i,j}$ and $\hat{X}_{i,j}$ are the values of the j^{th} ground-truth and prediction trajectories at time t_i, respectively:

$$\text{MSE}(X, \widehat{X}) = \sum_{i=1}^{N} \frac{1}{M} \sum_{j=1}^{M} (X_{i,j} - \hat{X}_{i,j})^2 \Delta t.$$

4. The sum of squared distance between mean trajectories and absolute distance between trajectories, which is a common practice for estimating the parameters of an SDE. Here M and $X_{i,j}$ and $\hat{X}_{i,j}$ have the same meaning as in the MSE definition. $\text{var}(X_i)$ and $\text{var}(\hat{X}_i)$ are the variances of the empirical distributions of $X(t_i), \hat{X}(t_i)$, respectively. We shall denote this loss function by

$$(\text{mean}^2 + \text{var})(X, \hat{X}) = \sum_{i=1}^{N} \left[\left(\frac{1}{M} \sum_{j=1}^{M} X_{i,j} - \frac{1}{M} \sum_{j=1}^{N} \hat{X}_{i,j} \right)^2 + \left| \text{var}(X_i) - \text{var}(\hat{X}_i) \right| \right] \Delta t.$$

5. Negative approximate log-likelihood of the trajectories:

$$-\log \mathcal{L}(X|\sigma) = -\sum_{i=0}^{N-1} \sum_{j=1}^{M} \log \rho_{\mathcal{N}} \left[\frac{X_{i+1,j} - X_{t,j} + f(X_{i,j}, t_i)\Delta t}{\sigma^2(X_{i,j}, t_i)\Delta t} \right],$$

where $\rho_{\mathcal{N}}$ is the probability density function of the standard normal distribution and $f(X_{i,j}, t_i), \sigma(X_{i,j}, t_i)$ are the ground-truth drift and diffusion functions in Eq. (1). M and $X_{i,j}$ and $\hat{X}_{i,j}$ have the same meaning as in the MSE definition.

6. MMD (maximum mean discrepancy) (Li et al., 2015):

$$\text{MMD}(X, \hat{X}) = \sum_{i=1}^{N} \Big(\mathbb{E}_p[K(X_i, X_i)] - 2\mathbb{E}_{p,q}[K(X_i, \hat{X}_i)] + \mathbb{E}_q[K(\hat{X}_i, \hat{X}_i)] \Big) \Delta t,$$

where K is the standard radial basis function (or Gaussian kernel) with multiplier 2 and number of kernels 5. X_i and $\hat{X}_i$ are the values of the ground-truth and prediction trajectories at time t_i, respectively.

Five W_2-distance-based loss functions for the 2D SDE reconstruction problem Example 4 are listed as follows

1. 2D squared W_2 loss

$$\sum_{i=1}^{N-1}\Big(W_2^2\big(\mu_{N,1}(t_i),\hat{\mu}_{N,1}(t_i)\big)+W_2^2\big(\mu_{N,2}(t_i),\hat{\mu}_{N,2}(t_i)\big)\Big)\Delta t$$

where $\mu_{N,1}(t_i)$ and $\hat{\mu}_{N,1}(t_i)$ are the empirical distributions of $X_1, \hat{X}_1$ at time t_i, respectively. Also, $\mu_{N,2}(t_i)$ and $\hat{\mu}_{N,2}(t_i)$ are the empirical distributions of $X_2, \hat{X}_2$ at time t_i, respectively.

2. Weighted sliced squared W_2 loss

$$\sum_{i=1}^{N-1}\Big(\sum_{k=1}^{m}\frac{N_k}{\sum_{\ell=1}^{m}N_\ell}W_2^2\big(\mu_{N,k}^{\mathrm{s}}(t_i),\hat{\mu}_{N,k}^{\mathrm{s}}(t_i)\big)\Big)\Delta t$$

where $\mu_{N,k}^{\mathrm{s}}(t_i)$ is the empirical distribution for $\sqrt{X_1(t_i)^2+X_2(t_i)^2}$ such that the angle between the two vectors $\big(X_1(t_i),X_2(t_i)\big)$ and (1, 0) is in $[\frac{2(k-1)\pi}{m},\frac{2k\pi}{m})$; $\hat{\mu}_{N,k}^{\mathrm{s}}(t_i)$ is the empirical distribution for $\sqrt{\hat{X}_1(t_i)^2+\hat{X}_2(t_i)^2}$ such that the angle between the two vectors $\big(\hat{X}_1(t_i),\hat{X}_2(t_i)\big)$ and (1, 0) is in $[\frac{2(k-1)\pi}{m},\frac{2k\pi}{m})$; N_k is the number of predictions such that the angle between the two vectors $(\hat{X}_1(t_i),\hat{X}_2(t_i))$ and (1, 0) is in $\big[\frac{2(k-1)\pi}{m},\frac{2k\pi}{m}\big)$.

3. The loss function Eq. (20)

$$W_2^2(\mu_N^{\mathrm{e}},\hat{\mu}_N^{\mathrm{e}}),$$

where μ_N^{e} and $\hat{\mu}_N^{\mathrm{e}}$ are the empirical distributions of the vector $(\boldsymbol{X}(t_1),...\boldsymbol{X}(t_{N-1}))$ and $(\hat{\boldsymbol{X}}(t_1),...,\hat{\boldsymbol{X}}(t_{N-1}))$, respectively. It is estimated by

$$W_2^2(\mu_N^{\mathrm{e}},\hat{\mu}_N^{\mathrm{e}})\approx\mathrm{ot.emd2}\Big(\frac{1}{M}\boldsymbol{I}_M,\frac{1}{M}\boldsymbol{I}_M,\boldsymbol{C}\Big),\tag{A38}$$

where $\mathrm{ot.emd2}$ is the function for solving the earth movers distance problem in the ot package of Python, M is the number of ground-truth and predicted trajectories, $\boldsymbol{I}_\ell$ is an M-dimensional vector whose elements are all 1, and $\boldsymbol{C}\in\mathbb{R}^{M\times M}$ is a matrix with entries $(\boldsymbol{C})_{ij}=|\boldsymbol{X}_N^i-\hat{\boldsymbol{X}}_N^j|_2^2$. $\boldsymbol{X}_N^i$ is the vector of the values of the i^{th} ground-truth trajectory at time points $t_1,...,t_{N-1}$, and $\hat{\boldsymbol{X}}_N^j$ is the vector of the values of the j^{th} predicted trajectory at time points $t_1,...,t_{N-1}$.

4. The time-decoupled squared W_2 loss function, which is the right-hand side of Eq. (29), estimated by

$$\sum_{i=1}^{N-1} \inf_{\pi_i} \mathbb{E}_{\pi_i}\big[|\boldsymbol{X}(t_i)-\hat{\boldsymbol{X}}(t_i)|_2^2\big]\Delta t \approx \sum_{i=1}^{N-1} W_2^2\big(\mu_N^{\mathrm{e}}(t_i),\hat{\mu}_N^{\mathrm{e}}(t_i)\big)\Delta t \approx \Delta t \sum_{i=1}^{N-1}\mathrm{ot.emd2}\Big(\frac{1}{M}\boldsymbol{I}_M,\frac{1}{M}\boldsymbol{I}_M,\boldsymbol{C}_i\Big), \tag{A39}$$

where $\mu_N^{\mathrm{e}}(t_i),\hat{\mu}_N^{\mathrm{e}}(t_i)$ are the empirical distribution of $\boldsymbol{X}(t_i)$, $\hat{\boldsymbol{X}}(t_i)$, respectively, and $\mathrm{ot.emd2}$ is the function for solving the earth movers distance problem in the ot package of Python. M is the number of ground-truth and predicted trajectories, and $\boldsymbol{I}_M$ is an ℓ -dimensional vector whose elements are all 1. Here, the matrix $\boldsymbol{C}_i\in\mathbb{R}^{M\times M}$ has entries $(\boldsymbol{C}_i)_{sj}=|\boldsymbol{X}^s(t_i)-\hat{\boldsymbol{X}}^j(t_i)|_2^2$ for $i=1,...,N-1$. $\boldsymbol{X}^s(t_i)$ is the vector of the values of the s^{th} ground-truth trajectory at the time point t_i, and $\hat{\boldsymbol{X}}^j(t_i)$ is the vector of the values of the j^{th} predicted trajectory at the time point t_i.

5. MMD (maximum mean discrepancy) (Li et al., 2015):

$$\mathrm{MMD}(\boldsymbol{X},\hat{\boldsymbol{X}})=\sum_{i=1}^{N}\Big(\mathbb{E}_p[K(\boldsymbol{X}_i,\boldsymbol{X}_i)]-2\mathbb{E}_{p,q}[K(\boldsymbol{X}_i,\hat{\boldsymbol{X}}_i)]+\mathbb{E}_q[K(\hat{\boldsymbol{X}}_i,\hat{\boldsymbol{X}}_i)]\Big)\Delta t,$$

where K is the standard radial basis function (or Gaussian kernel) with multiplier 2 and number of kernels 5. $\boldsymbol{X}_i$ and $\hat{\boldsymbol{X}}_i$ are the values of the ground-truth and prediction trajectories at time t_j, respectively.

Appendix F: Uncertainty in the initial condition

For reconstructing the CIR model Eq. (45) in Example 2, instead of using the same initial condition for all trajectories, we shall investigate the numerical performance of our proposed squared W_2 distance loss when the initial condition is not fixed, but rather sampled from a distribution.

First, we construct an additional dataset of the CIR model to allow the initial value $u_0\sim\mathcal{N}(2,\delta^2)$, with δ^2 ranging from 0 to 1, and $\mathcal{N}$ stands for the 1D normal distribution. We then train the model by minimizing Eq. (29) to reconstruct Eq. (45) with the same hyperparameters as in Example 2. The results are shown in Table 2, which indicate our proposed squared W_2 loss function is rather insensitive to the “noise", i.e., the variance in the distribution of the initial condition.

Table 2 Reconstructing the CIR model Eq. (45) when $u_0 \sim \mathcal{N}(2, \delta^2)$ with different variance δ^2

The results indicate that the reconstruction results are not sensitive to the variance in the distribution of the initial value u_0

Loss	δ	Relative Errors in f	Relative Errors in σ	N_{repeats}
W_2	0.0	0.072 (± 0.008)	0.071 (± 0.023)	10
W_2	0.1	0.053 (± 0.008)	0.043 (± 0.016)	10
W_2	0.2	0.099 (± 0.007)	0.056 (± 0.019)	10
W_2	0.3	0.070 (± 0.014)	0.083 (± 0.026)	10
W_2	0.4	0.070 (± 0.014)	0.078 (± 0.040)	10
W_2	0.5	0.075 (± 0.013)	0.138 (± 0.021)	10
W_2	0.6	0.037 (± 0.018)	0.069 (± 0.017)	10
W_2	0.7	0.075 (± 0.016)	0.043 (± 0.014)	10
W_2	0.8	0.041 (± 0.012)	0.079 (± 0.023)	10
W_2	0.9	0.082 (± 0.015)	0.108 (± 0.033)	10
W_2	1.0	0.058 (± 0.024)	0.049 (± 0.025)	10

Table 3 Reconstructing the CIR model when using neuron networks of different widths and numbers in each hidden layer to parameterize $\hat{f}, \hat{\sigma}$ in Eq. (2)

Loss	Width	Layer	Relative errors in f	Relative errors in σ	N_{repeats}
W_2	16	1	$0.131(\pm0.135)$	$0.170(\pm0.102)$	10
W_2	32	1	$0.041(\pm0.008)$	$0.109(\pm0.026)$	10
W_2	64	1	$0.040(\pm0.008)$	$0.104(\pm0.019)$	10
W_2	128	1	$0.040(\pm0.008)$	$0.118(\pm0.019)$	10
W_2	32	2	$0.049(\pm0.015)$	$0.123(\pm0.020)$	10
W_2	32	3	$0.094(\pm0.013)$	$0.166(\pm0.041)$	10
W_2	32	4	$0.124(\pm0.020)$	$0.185(\pm0.035)$	10
W_2	32	5	$0.041(\pm0.008)$	$0.122(\pm0.024)$	10
W_2	32	6	$0.043(\pm0.013)$	$0.117(\pm0.024)$	10
W_2	32	7	$0.044(\pm0.012)$	$0.109(\pm0.017)$	10

Appendix G: Neural network structure

We examine how the neural network structure affects the reconstruction of the CIR model Eq. (45) in Example 2. We vary the number of layers and the number of neurons in each layer (the number of neurons are set to be the same in each hidden layer), and the results are shown in Table 3.

The results in Table 3 show that increasing the number of neurons in each layer improves the reconstruction accuracy in σ. For reconstructing the CIR model in Example 2, using 32 neurons in each layer seems to be sufficient. On the other hand, when each layer contains 32 neurons, the number of hidden layers in the neural network seems does not affect the reconstruction accuracy of f, σ, and this indicates even 1 or 2 hidden layers are sufficient for the learning of f, σ. Thus, reconstructing the CIR model in Example 2 using our proposed squared W_2 based loss function does not require using complex deep or wide neural networks.

We also consider using the ResNet neural network structure (He et al., 2016). However, the application of the ResNet technique does not improve the reconstruction accuracy of the CIR model in Example 2. This is because simple feedforward multilayer neural network structure can work well for learning Eq. (45) when learning both f and σ so we do not need deep neural networks. Thus. the ResNet technique is not required. The results are shown in Table 4.

Appendix H: Using the stochastic gradient descent method for optimization

Here, we shall reconstruct the OU process Eq. (46) in Example 3 with the initial condition $X(0) = 0$ using the MMD and our squared W_2 distance loss functions Eqs. (20) and (29) with different numbers of ground-truth trajectories and different batch sizes for applying the stochastic gradient descent technique for optimizing the parameters in the neural networks for reconstructing the SDE.

We train 2000 epochs with a learning rate 0.001 for all numerical experiments. In all cases, the loss functions converge before 2000 epochs. From Table 5, for all three loss function, i.e., the MMD loss, Eqs. (20) and (29), a larger number of training samples leads to more accurate reconstruction of σ (the noise term). Furthermore, it can be seen from Table 5 that using a smaller batch size (16) for training tends to lead to less accurate reconstruction of σ for the MMD and Eq. (20) loss functions even if the number of trajectories in the training set is large. This feature might arise because the trajectories are intrinsically noisy and evaluating MMD and Eq. (20) will be inaccurate if the batch size is small. Therefore, using a smaller batch size does not remedy the high cost of MMD as the reconstruction error is large and leads to inaccurate reconstruction of the ground-truth SDE for smaller N_{sample}. On the other hand, our proposed time-decoupled squared W_2 distance loss function Eq. (29) gives similar performance in learning f, σ for both a batch size of 16 and a batch size of 256. In other words, using Eq. (29) is more robust to a smaller batch size. From Table 5, using a smaller batch size (16) leads to faster training. Thus, we can consider using Eq. (29) as the loss function together with a smaller batch size to boost training efficiency.

From the results in both Example 3 and Table 5, our proposed time-decoupled squared W_2 distance Eq. (29) is faster and more efficient than the MMD method and Eq. (20), making it potentially most suitable among all three loss functions for reconstructing SDEs.

Appendix I: Additional discussion on the loss functions Eqs. (20) and (29)

Here, we make an additional comparison between using Eqs. (20) and (29) as loss functions in Example 4. We set the number of training samples to be 128 and other hyperparameters for training to be the same as those in Example 4, as detailed in Table 1.

Table 4 Reconstructing the CIR model Eq. (45) when neuron networks have different numbers of hidden layers and are equipped with the ResNet technique

Loss	Layer	Relative errors in f	Relative errors in σ	N_{repeats}
W_2	1	0.045(±0.012)	0.116(±0.025)	10
W_2	2	0.053(±0.011)	0.108(±0.024	10
W_2	3	0.071(±0.017)	0.117(±0.040)	10
W_2	4	0.096(±0.035)	0.149(±0.064)	10

Each hidden layer contains 32 neurons

Table 5 Errors and runtime for different loss functions and different numbers of ground-truth trajectories when the training batch size is fixed to 16 and 256. The unit of runtime is hours

Loss	N_{sample}	Batch size	Relative error in f	Relative error in σ	Runtime	N_{repeats}
MMD	64	16	0.30 ± 0.12	0.49 ± 0.17	1.19 ± 0.59	10
MMD	128	16	0.30 ± 0.09	0.50 ± 0.20	1.27 ± 0.58	10
MMD	256	16	0.31 ± 0.09	0.44 ± 0.21	1.31 ± 0.59	10
MMD	512	16	0.22 ± 0.12	0.43 ± 0.18	1.22 ± 0.37	10
MMD	1024	16	0.23 ± 0.11	0.37 ± 0.24	1.70 ± 0.47	10
Eq. (20)	64	16	0.28 ± 0.06	0.66 ± 0.11	0.83 ± 0.26	10
Eq. (20)	128	16	0.24 ± 0.07	0.68 ±0.11	0.73 ± 0.18	10
Eq. (20)	256	16	0.25 ± 0.07	0.66 ± 0.09	0.67 ± 0.14	10
Eq. (20)	512	16	0.23 ± 0.06	0.68 ± 0.09	0.75 ± 0.16	10
Eq. (20)	1024	16	0.25 ± 0.07	0.66 ± 0.09	1.02 ± 0.47	10
Eq. (29)	64	16	0.20 ± 0.06	0.42 ± 0.08	0.61 ± 0.14	10
Eq. (29)	128	16	0.22 ± 0.06	0.37 ± 0.14	0.78 ± 0.35	10
Eq. (29)	256	16	0.21 ± 0.07	0.39 ± 0.16	0.88 ± 0.46	10
Eq. (29)	512	16	0.23 ± 0.06	0.43 ± 0.15	0.72 ± 0.11	10
Eq. (29)	1024	16	0.21 ± 0.03	0.36 ± 0.12	1.08 ± 0.52	10
MMD	64	256	0.26 ± 0.12	0.41 ± 0.20	1.54 ± 0.66	10
MMD	128	256	0.25 ± 0.14	0.40 ± 0.23	1.82 ± 0.78	10
MMD	256	256	0.25 ± 0.12	0.35 ± 0.21	3.68 ± 1.31	10
MMD	512	256	0.23 ± 0.14	0.37 ± 0.23	3.45 ± 1.50	10
MMD	1024	256	0.23 ± 0.13	0.35 ± 0.21	3.09 ± 1.35	10
Eq. (20)	64	256	0.28 ± 0.08	0.61 ± 0.04	1.19 ± 0.45	10
Eq. (20)	128	256	0.31 ± 0.07	0.61 ± 0.07	1.04 ± 0.48	10
Eq. (20)	256	256	0.26 ± 0.07	0.53 ± 0.03	0.96 ± 0.43	10
Eq. (20)	512	256	0.26 ± 0.08	0.56 ± 0.05	0.98 ± 0.40	10
Eq. (20)	1024	256	0.27 ± 0.08	0.56 ± 0.05	0.89 ± 0.36	10
Eq. (29)	64	256	0.24 ± 0.08	0.41 ± 0.13	1.39 ± 0.53	10
Eq. (29)	128	256	0.26 ± 0.11	0.37 ± 0.17	1.36 ± 0.61	10
Eq. (29)	256	256	0.20 ± 0.08	0.31 ± 0.16	1.72 ± 0.73	10
Eq. (29)	512	256	0.25 ± 0.11	0.38 ± 0.20	1.67 ± 0.73	10
Eq. (29)	1024	256	0.26 ± 0.10	0.39 ± 0.20	1.64 ± 0.79	10

The MMD and our proposed squared W_2 distance Eq. (20) and well as our proposed time-decoupled squared W_2 distance Eq. (29) are used as the loss function

First, we minimize Eq. (20) and record Eqs. (20) and (29) over training epochs. Next, we minimize Eq. (29) and record Eq. (20) and Eq. (29) over training epochs. The results are shown in Fig. 6.

From Fig. 6a, we can see that when minimizing Eq. (20, Eq. 20) is almost $10^{0.5}$ times larger than Eq. (29). However, when minimizing Eq. (29), the values of Eqs. (20) and (29) are close to each other (Fig. 6b). In both cases, Eq. (29) converges to approximately 10^{-1}. Interestingly, minimizing Eq. (29) leads to a smaller value of Eq. (20). This again implies that minimizing Eq. (29) can be more effective than minimizing Eq. (20) in Example 4. More analysis on Eq. (29) is needed to understand its theoretical properties and to compare the performances of minimizing Eq. (29) versus minimizing Eq. (20) from numerical aspects.

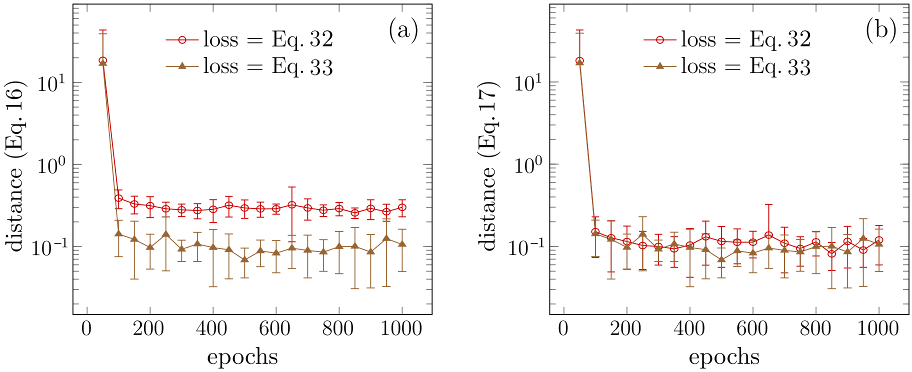

Fig. 6 **a** The change in Eqs. (20) and (29) when minimizing Eq. (20) over training epochs. **b** The change in Eqs. (20) and (29) when minimizing Eq. (29) over training epochs

Acknowledgements The authors acknowledge Dr. Hedi Xia for his constructive comments on this manuscript.

Author contributions MX and XL collaborated and designed the methodology. MX provided mathematical proof of the theoretical results. XL and QS ran the numerical experiments. MX, XL, and TC wrote the manuscript. All authors reviewed and revised the manuscript.

Funding Tom Chou acknowledges support from the Army Research Office through grant W911NF-18-1-0345.

Data availability No data was used during this study. All codes will be publicly available upon acceptance of this manuscript.

Material availability Code that generates the data will be made publicly available on Github upon acceptance of this manuscript.

Code availability All training details have been specificied in Table 1. Codes for reproducing numerical results in this manuscript are available at https://github.com/mtxia99/learning_neural_sde.

Declarations

Conflict of interest No competing and financial interests to disclose.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In: *International conference on machine learning*, pp. 214–223. PMLR

Bartl, D., Beiglböck, M., & Pammer, G. (2021). The Wasserstein space of stochastic processes. arXiv preprint arXiv:2104.14245

Bion-Nadal, J., & Talay, D. (2019). On a Wasserstein-type distance between solutions to stochastic differential equations. *The Annals of Applied Probability, 29*(3), 1609–1639. https://doi.org/10.1214/18-aap1423

Bressloff, P. C. (2014). *Stochastic Processes in Cell Biology* (Vol. 41). Springer.

Briol, F.-X., Barp, A., Duncan, A.B., & Girolami, M. (2019). Statistical inference for generative models with maximum mean discrepancy. arXiv preprint arXiv:1906.05944

Chen, R.T., Rubanova, Y., Bettencourt, J., & Duvenaud, D.K. (2018). Neural ordinary differential equations. In: *Advances in neural information processing systems*, vol. 31

Chewi, S., Clancy, J., Le Gouic, T., Rigollet, P., Stepaniants, G., & Stromme, A. (2021). Fast and smooth interpolation on Wasserstein space. In: *International conference on artificial intelligence and statistics*, pp. 3061–3069. PMLR

Choulli, T., & Schweizer, M. (2024). New stochastic F ubini theorems. arXiv preprint https://arxiv.org/abs/2403.13791

Cinlar, E. (2011). *Probability and Stochastics*. Springer.

Clement, P., & Desch, W. (2008). An elementary proof of the triangle inequality for the Wasserstein metric. *Proceedings of the American Mathematical Society, 136*(1), 333–339.

Cuturi, M., Teboul, O., & Vert, J.-P. (2019). Differentiable ranks and sorting using optimal transport. In: *Proceedings of the 33rd international conference on neural information processing systems*, pp. 6861–6871

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). Pot: python optimal transport. *The Journal of Machine Learning Research, 22*(1), 3571–3578.

Fournier, N., & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields, 162*(3–4), 707–738.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., & Poggio, T.A. (2015). Learning with a Wasserstein loss. In: *Advances in neural information processing systems*, vol. 28

Goldbeter, A. (1995). A model for circadian oscillations in the Drosophila period protein (PER). *Proceedings of the Royal Society of London Series B: Biological Sciences, 261*(1362), 319–324.

Gonze, D. (2011). Modeling circadian clocks: From equations to oscillations. *Central European Journal of Biology, 6*, 699–711.

Gzyl, H., Horst, E., & Malone, S. W. (2008). Bayesian parameter inference for models of the black and Scholes type. *Applied Stochastic Models in Business and Industry, 24*(6), 507–524.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778

Jacod, J. (2006). *Calcul stochastique et problemes de martingales* (Vol. 714). Springer.

Jia, J., & Benson, A.R. (2019). Neural jump stochastic differential equations. In: *Advances in neural information processing systems32*

Kidger, P., Foster, J., Li, X., & Lyons, T.J. (2021). Neural SDEs as infinite-dimensional GANs. In: *International conference on machine learning*, pp. 5453–5463. PMLR

Kloeden, P. E., & Platen, E. (1992). *Numerical solution of stochastic differential equations*. Springer.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., & Rohde, G. (2019). Generalized sliced Wasserstein distances. In: *Advances in neural information processing systems*, vol. 32

Kolouri, S., Rohde, G.K., & Hoffmann, H. (2018). Sliced Wasserstein distance for learning Gaussian mixture models. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3427–3436

Li, Y., Swersky, K., & Zemel, R. (2015). Generative moment matching networks. In: *International conference on machine learning*, pp. 1718–1727. PMLR

Li, X., Wong, T.-K.L., Chen, R.T., & Duvenaud, D. (2020). Scalable gradients for stochastic differential equations. In: *International conference on artificial intelligence and statistics*, pp. 3870–3882. PMLR

Lin, Y. T., & Buchler, N. E. (2018). Efficient analysis of stochastic gene dynamics in the non-adiabatic regime using piecewise deterministic Markov processes. *Journal of The Royal Society Interface, 15*(138), 20170804.

Liu, H., Ong, Y.-S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: a review of scalable GPs . *IEEE Transactions on Neural Networks and Learning Systems, 31*(11), 4405–4423.

MacKay, D. J., et al. (1998). Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences, 168*, 133–166.

Mathelin, L., Hussaini, M. Y., & Zang, T. A. (2005). Stochastic approaches to uncertainty quantification in cfd simulations. *Numerical Algorithms, 38*, 209–236.

Musiela, M., & Rutkowski, M. (2006). *Martingale methods in financial modelling* vol. 36. Springer

Oh, J.H., Pouryahya, M., Iyer, A., Apte, A.P., Tannenbaum, A., & Deasy, J.O. (2019). Kernel Wasserstein distance. arXiv preprint arXiv:1905.09314

Pereira, J., Ibrahimi, M., & Montanari, A. (2010). Learning networks of stochastic differential equations. In: *Advances in neural information processing systems*, vol. 23

Rowland, M., Hron, J., Tang, Y., Choromanski, K., Sarlos, T., & Weller, A. (2019). Orthogonal estimation of Wasserstein distances. In: *The 22nd International conference on artificial intelligence and statistics*, pp. 186–195. PMLR

Rüschendorf, L. (1985). The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields, 70*(1), 117–129.

Sanz-Serna, J. M., & Zygalakis, K. C. (2021). Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *The Journal of Machine Learning Research, 22*(1), 11006–11042.

Soize, C. (2017). *Uncertainty quantification*. Springer.

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. In: *International conference on learning representations*

Sullivan, T. J. (2015). *Introduction to uncertainty quantification* (Vol. 63). Springer.

Tong, A., Nguyen-Tang, T., Tran, T., & Choi, J. (2022). Learning fractional white noises in neural stochastic differential equations. *Advances in Neural Information Processing Systems, 35*, 37660–37675.

Tzen, B., & Raginsky, M. (2019). Neural stochastic differential equations: deep latent Gaussian models in the diffusion limit. arXiv preprint arXiv:1905.09883

Vecchi, F. C., Morando, P., & Ugolini, S. (2016). Reduction and reconstruction of stochastic differential equations via symmetries. *Journal of Mathematical Physics, 57*(12), Article 123508.

Villani, C., et al. (2009). *Optimal transport: Old and new* (Vol. 338). Springer.

Wang, J. (2016). L^p-Wasserstein distance for stochastic differential equations driven by Lévy processes. *Bernoulli*, pp. 1598–1616

Welch, G.F. (2020). Kalman filter. *Computer vision: A reference guide*, pp. 1–3

Welch, G., Bishop, G., et al. (1995). An introduction to the Kalman filter

Westermark, P. O., Welsh, D. K., Okamura, H., & Herzel, H. (2009). Quantification of circadian rhythms in single cells. *PLOS Computational Biology, 5*(11), 1–10. https://doi.org/10.1371/journal.pcbi.1000580

Xia, M., Li, X., Shen, Q., & Chou, T. (2024). *An efficient Wasserstein-distance approach for reconstructing jump-diffusion processes using parameterized neural networks*. Machine Learning: Science and Technology, 5(4), pp. 045052.

Zheng, W., Wang, F.-Y., & Gou, C. (2020). Nonparametric different-feature selection using Wasserstein distance. In: *2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI)*, pp. 982–988. IEEE

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