Nonparametric inference of stochastic differential equations based on the relative entropy rate

Min Dai¹ | Jinqiao Duan² | Jianyu Hu³ | Xiangjun Wang³

¹School of Science, Wuhan University of Technology, Wuhan, China
²Department of Applied Mathematics, College of Computing, Illinois Institute of Technology, Chicago, Illinois, USA
³School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China

Correspondence
Jianyu Hu, School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China. Email: jyhu94@outlook.com

Communicated by: I. Stratis

Funding information
National Natural Science Foundation of China, Grant/Award Number: 12001213, 11801192 and 11771449; National Science Foundation (NSF), Grant/Award Number: 1620449

The information detection of complex systems from data is currently undergoing a revolution, driven by the emergence of big data and machine learning methodology. Discovering governing equations and quantifying the dynamical properties of complex systems are among the central challenges. In this work, we devised a nonparametric approach to learning the relative entropy rate from observations of stochastic differential equations with different drift functions. The estimator corresponding to the relative entropy rate is then presented via the Gaussian process kernel theory. Meanwhile, this approach enables us to extract the governing equations. We illustrate our approach with several examples. Numerical experiments show the proposed approach performs well for rational drift functions, not only polynomial drift functions.

KEYWORDS
Gaussian process kernel theory, nonparametric approach, relative entropy rate, stochastic differential equations

MSC CLASSIFICATION
60G15, 60H10, 60H35

1 | INTRODUCTION

Dynamical properties of complex systems such as complexity and sensitivity are ubiquitous, which are significant in the research of stochastic dynamical systems.¹–³ It is less well understood, however, how to better quantify the dynamical properties of stochastic dynamics. The relative entropy rate is an effective tool to characterize the complexity and sensitivity of dynamical properties.⁴,⁵ Traditional methods are affected by factors such as unknown governing equations. Owing to the declining cost of data storage and computation, as well as the development of machine learning, data-driven discovery methodologies have made great progress. Combining with data-driven discovery methodologies, it is possible to compute the relative entropy rate from time series data.

There exist many different forms in regard to data-driven methods, such as parametric and nonparametric approaches.⁶–¹⁰ The recent sparse identification of the nonlinear dynamics method, which was proposed by Brunton and co-workers,¹¹ is a scripture of parametric approaches. Opper¹² employs the parametric techniques to compute the relative entropy rate. Whereas some complex systems are too complicated to model precisely via parametric representation, the nonparametric techniques, for example, neural networks¹³,¹⁴ and Gaussian processes,¹⁵,¹⁶ thus are applied in data-driven modeling. They outperform the parametric methods for obtaining the data-driven models with accuracy and extensiveness, although the parametric methods provide a concrete manifestation of models.
The relative entropy or Kullback–Leibler divergence has recently reemerged in machine learning as a cost function describing the difference between two probability distributions.\textsuperscript{17,18} It provides a precise characterization for approximating the underlying probability distribution by the probability distribution from data. The Kullback–Leibler divergence has been applied to infinite dimensional probability measures of stochastic processes, such as probability inference\textsuperscript{19} and optimal control problems.\textsuperscript{20–22} The relative entropy rate,\textsuperscript{23} which is the long-term average limit of the Kullback-Leibler divergence, is a distance measure between two stochastic processes. It plays an important role in quantifying dynamical properties. Dupuis et al\textsuperscript{4} discussed the uncertainty and sensitivity bounds of stochastic dynamics from observation data using the relative entropy rate. Pantazis and Katsoulakis\textsuperscript{5} applied the relative entropy rate as a suitable information-theoretic object to analyze the sensitivity of the probability distribution of stochastic processes in regard to perturbations in the parameters of the complex dynamics. Techakesari et al\textsuperscript{24} use the relative entropy rate to design hybrid system filters in the presence of (parameterized) model uncertainty.

In this work, we propose a nonparametric approach to learning the relative entropy rate from observations of stochastic differential equations with different drift functions. Current efforts are mainly focused on deriving a simple estimator for relative entropy rate via the Gaussian process kernel theory. Meanwhile, this approach enables us to extract the governing information-theoretic object to analyze the sensitivity of the probability distribution of stochastic processes in regard to perturbations in the parameters of the complex dynamics. The remainder of this paper is structured as follows. In Section 2, we introduce the relative entropy rate of stochastic differential equations and the variational formulation. In Section 3, we propose a nonparametric technique for the calculation of the relative entropy rate based on one sample of stochastic differential equations. Some numerical examples are presented in Section 4, followed by a conclusion in Section 5.

2 THE RELATIVE ENTROPY RATE

In this work, we consider a stochastic differential equation for the dynamics of a \(d\)-dimensional diffusion process \(X_t\), given by

\[
dX_t = g(X_t)dt + \sigma(X_t)dB_t,
\]

with initial data \(X_0 = x_0 \in \mathbb{R}^d\), where \(g\) is the drift function, \(\sigma\) is the \(d \times k\) dimensional matrix and \(B_t\) is the standard Brownian motion in \(\mathbb{R}^k\).

The generator \(A\) of this stochastic differential equation (1) is\textsuperscript{25}

\[
A\phi = g \cdot \nabla \phi + \frac{1}{2} \text{tr}[\sigma \sigma^T H(\phi)],
\]

for all \(\phi\) in Sobolev space \(H_0^2(\mathbb{R}^d)\), where \(H\) is the Hessian operator with \(H = \nabla \nabla^T\). Thus, the probability density \(p(x, t)\) of the solution process \(X_t\) satisfies the Fokker–Planck equation

\[
\frac{\partial}{\partial t} p(x, t) = A^* p(x, t),
\]

with initial condition \(p(x, 0) = \delta(x - x_0)\), \(A^*\) is the adjoint operator of the generator \(A\) in Hilbert space \(L^2(\mathbb{R}^d)\), given by

\[
A^* \phi = -\nabla \cdot (g \phi) + \frac{1}{2} \text{tr}[H(\sigma \sigma^T \phi)].
\]

Here, \(H(\sigma \sigma^T \phi)\) is interpreted as matrix multiplication of \(H = \nabla \nabla^T\) and \(\sigma \sigma^T \phi\) (note that \(\phi\) is a scalar function). We call \(p\) the stationary probability density of the solution process \(X_t\) if it satisfies the stationary Fokker–Planck equation \(A^* p = 0\).

2.1 The relative entropy rate for stochastic differential equations

The relative entropy or Kullback–Leibler divergence between the probability measures \(P^g\) and \(P^r\) of two solution processes for stochastic differential equation (1) with different drifts \(g\) and \(r\) is defined as, see Appendix A for details\textsuperscript{12,17}

\[
E_{P^g} \left[ \ln \frac{P^g}{P^r} \right] = \frac{1}{2} \int_0^T dt \int p^g(x, t) ||g(x) - r(x)||_{\mathbb{R}^d}^2 dx.
\]
Here, the diffusion matrix of two solution processes are both $D(x) = \sigma(x)\sigma(x)^T$, $p^g(x,t)$ is the probability density of solution process with drift function $g$, and $||u(x)||_{A}^{2} = u(x) \cdot A(x)u(x)$ for some positive definite matrix $A$. Suppose that the stationary probability exists. Then, the probability density $p^g(x,t)$ converges to the stationary probability density $p^\psi$ as $t \to \infty$. Hence, consider the relative entropy rate as follows:

$$d(p^g, p^\psi) = \lim_{T \to \infty} \frac{1}{T} E_p \left[ \ln \frac{p^g}{p^\psi} \right] = \frac{1}{2} \int p^\psi(x)||g(x) - r(x)||_{D^{-1}}^2 dx. \quad (6)$$

In our work, we suppose that the drift $r$ and the diffusion $D$ are known, but the expression form of the drift $g$ is unknown. We, however, want to estimate the relative entropy rate via the observation data of the process $X_t$ with the drift $g$.

Based on Equation (6), the estimator of the drift $g$ and the stationary probability density $p^\psi$ are crucial to the calculation of the relative entropy rate. In order to simplify the estimation problem, we assume that the diffusion $D = \sigma^2 I$ and the drift $g$ satisfy a potential condition such as $g(x) = \nabla \psi(x)$. Then the stationary probability density fulfills $p^\psi(x) \propto e^{\frac{\psi(x)}{\sigma^2}}$, and one can estimate the density from observation data.

We will introduce a different way to calculate the relative entropy rate based on the variational formulation and generalized potential conditions. Inspired by a previous study,\textsuperscript{12} we also suppose that the expression of drift $g$ is

$$g(x) = r(x) + D(x)\nabla \psi^*(x). \quad (7)$$

Then the relative entropy rate (6) becomes

$$d(p^g, p^\psi) = \frac{1}{2} \int p^\psi(x)||\nabla \psi^*(x)||_{D}^2 dx. \quad (8)$$

Moreover, the stationary Fokker–Planck equation for drift $g$ is defined as

$$A^*_g p^\psi(x) = A^*_g p^g(x) - \nabla \cdot (D(x)\nabla \psi^*(x)p^\psi(x)) = 0, \quad (9)$$

where $p^\psi$ is the stationary probability density, and the operator $A^*_g$, corresponding to known drift $r$, can express as

$$A^*_g p^\psi(x) = -\nabla \cdot (r(x)p^\psi(x)) + \frac{1}{2} \text{tr}[\nabla \nabla^T(D(x)p^\psi(x))]. \quad (10)$$

### 2.2 Variational formulation

In this section, we will give a brief introduction to the variational formulation. Here, we assume that the specific form of the drift $g$ is unknown. We first estimate the drift $g$ from the observation data using the variational formulation for the stationary Fokker–Planck eq. 9 through the relative entropy rate. Suppose that the stationary probability density $p^\psi$ is given, and we search for an estimator of the drift $g$ by minimizing the relative entropy rate (6). Introducing a Lagrange multiplier function $\psi$, we may derive the drift $g$ from the following Lagrange functional

$$\frac{1}{2} \int p^\psi(x)||g(x) - r(x)||_{D^{-1}}^2 dx - \int \psi(x)A^*_g p^\psi(x) dx$$

$$= \frac{1}{2} \int p^\psi(x)||g(x) - r(x)||_{D^{-1}}^2 dx - \int \psi(x)\{A^*_g p^\psi(x) - \nabla \cdot ((g(x) - r(x))p^\psi(x))\} dx \quad (11)$$

The Fokker–Planck operator $A^*_g$ is defined in (9) for the drift $g$, and the operator $A^*_r$ is in (10).

Furthermore, making a variation of Lagrange functional (11) with respect to $g - r$, one can obtain $g(x) - r(x) = D(x)\nabla \psi(x)$. Inserting this result back into (11), the variational representation of relative entropy rate for a unknown potential $\psi$ is

$$\varepsilon_g[\psi] = \int \left\{ \frac{1}{2} ||\nabla \psi(x)||_{D}^2 + A_t \psi(x) \right\} p^\psi(x) dx, \quad (12)$$
where the generator $A_r$ is adjoint operator of $A^*_r$, and Equation (10) satisfies $\int \psi(x)A^*rp(x)dx = \int p(x)A\psi(x)dx$. In addition, the expression of the generator $A_r$ is

$$A_r\psi(x) = r(x) \cdot \nabla \psi(x) + \frac{1}{2}tr[D(x)\nabla^T \psi(x)].$$

We next introduce the variational bound\(^{12}\) for the Lagrange functional (12)

$$-\varepsilon_g[\psi] \leq \frac{1}{2} \int p^g(x) ||\nabla \psi^*(x)||_2^2 dx. \quad (13)$$

The equality is achieved when $\psi = \psi^*$. We are surprised to discover that, from the perspective of variational bound, the minimization of variational representation (12) can give us an estimator for the potential $\psi$ and also compute the relative entropy rate

$$\psi^*(x) = \arg \min_{\psi(x)} \varepsilon_g[\psi(x)],$$

$$d(P^g, P^r) = -\varepsilon_g[\psi^*(x)]. \quad (14)$$

According to the form of Lagrange functional $\varepsilon_g[\psi]$, we know that by applying the observation data that are the ergodic samples of the process with the drift $g$, one can estimate every potential $\psi$ under the stationary probability density condition. The estimator of the relative entropy rate, then, can be obtained. In the next section, we will present a nonparametric estimator for the potential $\psi$ and the relative entropy rate.

### 3 | NONPARAMETRIC INFERENCE FOR RELATIVE ENTROPY RATE

In this framework, our goal is to construct a nonparametric estimator for the relative entropy rate. We however know from the above theory of the potential $\psi$ and the stationary probability density $p^g$ are essential to calculating the relative entropy rate. Hence, we first need a smooth estimator for stationary probability density $p^g$ and then compute the potential $\psi$ from data using a nonparametric approach via the variational representation (12).

There are many density estimation methods, among which kernel density estimation is commonly used. The drawback of the kernel density estimation is nevertheless that it neglects the temporal ordering of observation data, as well as the time scale between them. Here, we employ the empirical distribution to replace the exact probability $p^g$.\(^{15}\)

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i). \quad (15)$$

The data $x_1, x_2, \cdots, x_n$ are random, ergodic samples of the process with the exact probability $p^g$. Moreover, one usually builds a parametric estimator of the potential $\psi$, which is represented by a finite set of basis functions. However, this representation of the parametric method is not sufficient to express many functions, such as rational functions. We, therefore, work on a nonparametric estimate, which is a more general approach to these functions.

We now demonstrate the nonparametric estimate method. Introduce a penalty term firstly, which is selected as a quadratic form $\frac{1}{2} \sum_k \omega_k^2 / \lambda_k$, where the $\omega_k$ are the weights corresponding to the basic functions of the parametric representation and the $\lambda_k$ are hyper-parameters, to regularize the potential estimator due to a limited amount of observational data. We can also consider the penalty term from the perspective of a pseudo-Bayesian. The $\exp \left\{ -\frac{1}{2} \sum_k \omega_k^2 / \lambda_k \right\}$ is viewed as a Gaussian prior distribution over weights $\omega_k$. Assume that the parametric of the variational representation (12) is $\varepsilon_{\text{emp}}[\psi_{\omega}]$. At the same time, the $\exp \left\{ -C \varepsilon_{\text{emp}}[\psi_{\omega}] \right\}$ can be interpreted as a likelihood, where $C$ represents the proportionality weight between the penalty term and the observation data. The potential function $\psi$ can thus be treated as a Gaussian process according to the above interpretation. Inspired by the viewpoint of the Gaussian process, we will convert the parameter representation form into the kernel function form using the kernel trick, which makes the information for $\psi$ expressed fully. To this end, define

$$K(x, x') = \sum_k \lambda_k \phi_k(x) \phi_k(x'). \quad (16)$$
where $\lambda_k$ and $\phi_k$ are the orthonormal eigenvalues and eigenfunctions, respectively. This is also the covariance kernel of the prior of a Gaussian process about the potentials $\psi$.

The regularized functional of the potential function $\psi$ can be defined as

$$C \sum_{i=1}^{n} \left\{ \frac{1}{2} || \nabla \psi(x_i) ||_D^2 + A_r \psi(x_i) \right\} + \frac{1}{2} \int \int \psi(x)K^{-1}(x,x')\psi(x')dx'dx'. \quad (17)$$

via the kernel approach. The $K^{-1}(x,x')$ represents the inverse of the kernel operator. In addition, the penalty term in (17) can be proved to be equivalent to the reproducing kernel Hilbert space norm of the drifts $\psi$ defined by the kernel $K$. We will further derive a specific representation of the drift function estimator.

The variation of (17) with respect to $\psi$ yields

$$Cn[A_r^\dagger \hat{p}(x) - \nabla \cdot (D(x)\nabla \psi(x)\hat{p}(x))] + \int \int K^{-1}(x,x')\psi(x')dx' = 0.$$ 

Multiplying both sides by the operator $K$, one obtains

$$\psi(x) + C \sum_{j=1}^{n} ((A_g[\psi])_{x'=x_j}K(x,x'))_{x'=x_j} = 0. \quad (18)$$

where the generator acts on the kernel function $K$ as

$$(A_g[\psi]_{x'}K(x,x'))_{x'=x_j} = (r(x') + D(x')\nabla \psi(x'))\nabla K(x,x') + \frac{1}{2} tr[D(x')\nabla \nabla^T K(x,x')]. \quad (19)$$

From above Equations (18) and (19), we see that if we know $\nabla \psi(x)$ at all observation data $x = x_i$, the potential function value $\psi(x)$ can be calculated for all $x$. The key to this work, however, is to evaluate the gradient of the potential function. We next compute $\nabla \psi(x)$ at all observation data points via performing the gradient of Equation (18) and setting $x = x_i$. Thus, there are a series of linear equations

$$\nabla \psi(x_i) + C \sum_{j=1}^{n} (A_g[\psi])_{x'=x_j} \nabla_x K(x,x')_{x=x_i,x'=x_j} = 0. \quad (20)$$

We so far obtain the drift function value $\nabla \psi^*(x_i)$ at every point. Further, substituting into these drift function values, we can calculate the relative entropy rate

$$d(P^g, P^r) = \frac{1}{2n} \sum_{i=1}^{n} || \nabla \psi^*(x_i) ||_D^2 \quad (21).$$

### 4 NUMERICAL EXPERIMENTS

We begin our verification of the nonparametric estimate approach by describing a few experimental examples. Without special emphasis, we apply the radial basis function kernel

$$K(x,y) = \exp\left(-\frac{(x-y)^T(x-y)}{2l^2}\right). \quad (22)$$

where the length scale $l$ is the hyper-parameter. In this section, we also show the nonparametric learning results of the drift function in order to further illustrate the advantages of this method.

**Example 1.** Consider a scalar stochastic dynamical system with the polynomial drift term

$$dX_t = (4X_t - 4X_t^3 - \beta(X_t^2 + 2X_t + 1))dt + \sigma(X_t)dW_t, \quad X_0 = x. \quad (23)$$
where $W_t$ is the standard Brownian motion and the diffusion function is $\sigma = 1$. The known drift term $r(x) = 4x - 4x^3$ corresponds to a stochastic double-well system. Another drift function, however, which generates the samples through long-term observation is $g_\beta(x) = 4x - 4x^3 + \nabla \psi(x)$, where $\nabla \psi(x) = -\beta(x^2 + 2x + 1)$ depends on $\beta$. The drifts $g$ and $r$ are equal when $\beta = 0$.

In the numerical simulation, we take the time step $\delta t = 0.001$ and use the Euler scheme to sample the stochastic differential equation with drifts $g_\beta$ regarding the different values of parameter $\beta$. Here, we apply $N = 10,000$ data points.

FIGURE 1 The nonparametric inference results of the drift function $\nabla \psi(x)$. (A) $\beta = 1$. (B) $\beta = 2$. (C) $\beta = 3$ [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2 The exact relative entropy rate and estimators corresponding to different $\beta$ for the stochastic double-well system [Colour figure can be viewed at wileyonlinelibrary.com]
$x_k = X_{t_k}$ are uniformly sampled from a trajectory with the observation time length $T = 10^5$. Thus, the discrete times are $t_k = k\tau$ for $k = 1, \ldots, N$ with $\tau = 10$. Employing these data points, we can estimate the potential function $\psi$ and its gradient, and further compute the relative entropy rate in terms of (14). The numerical results are presented in Figures 1 and 2. In Figure 1, we compare the accuracy and the learning results with regard to potential function $\psi$ for some $\beta$ values and discover that the learning results have a good performance. As shown in Figure 2, the exact relative entropy rate and estimators corresponding to different values of parameter $\beta$ are plotted. We can see that the exact relative entropy rate agrees well with the estimated result directly obtained from the nonparametric method.

**FIGURE 3** The nonparametric inference results of the drift function $\nabla \psi(x)$. (A) $\beta = 1$. (B) $\beta = 2$. (C) $\beta = 3$ [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 4** The exact relative entropy rate and estimators corresponding to different $\beta$ for the stochastic gene regulation system [Colour figure can be viewed at wileyonlinelibrary.com]
Example 2. Consider the transcription factor activator (TF-A) monomer concentration of stochastic differential equation in the gene regulation system

\[
dX_t = \beta \left( \frac{6X_t^2}{X_t^2 + 10} - X_t + 0.4 \right) \, dt + \sigma(X_t) \, dW_t, \quad X_0 = x, \tag{24}
\]

FIGURE 5 The nonparametric inference results of the drift function $\nabla \psi(x)$. (A) $\beta = 1.5$. (B) $\beta = 2$. (C) $\beta = 2.5$ [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 6 The exact relative entropy rate and estimators corresponding to different $\beta$ for the stochastic gene regulation system [Colour figure can be viewed at wileyonlinelibrary.com]
where the diffusion function is $\sigma(x) = 1$ and $W_t$ is the standard Brownian motion. The known drift term here is $r(x) = 0$, and another drift which generates the samples through long-term observation is $g_\beta(x) = \nabla \psi(x) = \beta((6x^2/(x^2 + 10)) - x + 0.4)$ depending on $\beta$. If parameter $\beta = 0$, the drifts $g$ and $r$ then are equal.

To further validate our method, taking the time step $\delta t = 0.001$ in the numerical simulation, and using the Euler scheme to sample the stochastic differential equation with drifts $g_\beta$ regarding the different values of parameter $\beta$, we generate a trajectory with the observation time length $T = 10^5$. Here, we utilize $N = 10,000$ data points $x_k = X_{tk}$, uniformly sampled from this generated trajectory. Thus, the discrete times are $t_k = kr, k = 1, \ldots, N$ with $r = 10$. Employing these data points, we can obtain the potential function $\psi$ and its gradient and, meanwhile, compute the relative entropy rate in terms of (14).

As shown in Figures 3 and 4, the comparisons regarding the learning and accuracy results of the potential function $\psi$ are given, and the exact relative entropy rate and estimators for different values of parameter $\beta$ are plotted. We can see that the evaluation of the potential function $\psi$ is pretty, and the exact relative entropy rate agrees well with the estimated result directly obtained from the nonparametric method.

**Example 3.** In order to show the superiority of our method, we also consider the trigonometric polynomial function

$$dX_t = \beta(\sin(X_t) - \sin(X_t)^3)dt + \sigma(X_t)dW_t, \ X_0 = x.$$  \hspace{1cm} (25)

Here, $W_t$ is the standard Brownian motion and the diffusion function $\sigma(x) = 1$. Taking the known drift term $r(x) = 0$, and another drift that generates the samples through long-term observation $g_\beta(x) = \nabla \psi(x) = \beta(\sin(X_t) - \sin(X_t)^3)$ depending on $\beta$. If parameter $\beta = 0$, the drifts $g$ and $r$ then are equal.

We take the time step $\delta t = 0.001$ in the numerical simulation and then apply the Euler scheme to sample the stochastic differential equation with drifts $g_\beta$ regarding the different values of parameter $\beta$ and generate a trajectory with the observation time length $T = 10^5$. Here, we utilize $N = 10,000$ data points $x_k = X_{tk}$, uniformly sampled from this generated trajectory, so that the discrete times are $t_k = kr, k = 1, \ldots, N$ with $r = 10$. Employing these data points, we obtain the potential function $\psi$ and its gradient and compute the relative entropy rate in terms of (14). We show the results in Figures 5 and 6.

The comparisons regarding the learning and accuracy results of the potential function $\psi$ are given in Figure 5, and the exact relative entropy rate and estimators for different values of parameter $\beta$ are plotted in Figure 6. We see that the evaluation of the potential function $\psi$ is pretty, and the exact relative entropy rate agrees well with the estimated result.

5 | CONCLUSION AND DISCUSSION

In conclusion, we have presented a nonparametric technique to learn the relative entropy rate, which is applied to describe dynamical properties. The study of the dynamical properties of systems is significant for stochastic dynamical systems. Here, we offer a good methodology for their study to learn from time series data. Moreover, the nonparametric method not only provides a precise estimation of our functions and the relative entropy rate but also adapts to a wider range of function types, such as rational functions, compared with the parametric approach. This performance has been well demonstrated in examples.

In addition, this work motivates a number of future extensions. Our framework can be extended in other Markov processes, including Markov chains or Lévy processes. They would be interested in a realistic application. On the other hand, the stationary density is worthy of attention because of the need for a long time of observation. We can generalize this issue with a finite time window $T$ and evaluate further the marginal densities $p^\delta(x, t)$.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) grants 12001213, 11801192, and 11771449 and National Science Foundation (NSF) grant 1620449.

CONFLICTS OF INTEREST

This work does not have any conflicts of interest.

This work does not have any conflicts of interest.
REFERENCES

1. Zhang Z, Ma X, Yu H, Hua H. Stochastic dynamics and sensitivity analysis of a multistage marine shafting system with uncertainties. *Ocean Eng*. 2021;219:108388.

2. Pantazis Y, Katsoulakis MA, Vlachos DG. Parametric sensitivity analysis for biochemical reaction networks based on pathwise information theory. *BMC Bioinform*. 2013;14:311-311.

3. Plyasunov S, Arkin AP. Efficient stochastic sensitivity analysis of discrete event systems. *J Comput Phys*. 2007;221:724-738.

4. Dupuis P, Katsoulakis MA, Pantazis Y, Plecháč P. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA J Uncertain Quant*. 2016;4(1):80-111.

5. Pantazis Y, Katsoulakis MA. A relative entropy rate method for path space sensitivity analysis of stationary complex stochastic dynamics. *J Chem Phys*. 2013;138(5):54115.

6. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. Neural ordinary differential equations. arXiv preprint arXiv:180607366; 2018.

7. Dai M, Gao T, Lu Y, Zheng Y, Duan J. Detecting the maximum likelihood transition path from data of stochastic dynamical systems. *Chaos*. 2020;30 11:113124.

8. Lei H, Baker NA, Li X. Data-driven parameterization of the generalized Langevin equation. *Proc Natl Acad Sci*. 2016;113(50):14183-14188.

9. Rudy SH, Brunton SL, Proctor JL, Kutz JN. Data-driven discovery of partial differential equations. *Sci Adv*. 2017;3(4):e1602614.

10. Batz P, Ruttor A, Opper M. Approximate gaussian process inference for the drift function in stochastic differential equations Advances in neural information processing systems; 2013:2040-2048.

11. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc Natl Acad Sci*. 2016;113(15):3932-3937.

12. Oppe M. An estimator for the relative entropy rate of path measures for stochastic differential equations. *J Stat Mech: Theory Exper*. 2016;2016(11):P11011.

13. Han J, Jentzen A, Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proc Natl Acad Sci*. 2018;115:8505-8510.

14. Ruttor A, Haber E. Deep neural networks motivated by partial differential equations. *J Math Imaging Vision*. 2019;62:352-364.

15. Batz P, Ruttor A, Oppe M. Variational estimation of the drift for stochastic differential equations from the empirical density. *J Stat Mech: Theory Exper*. 2016;2016(8):83404.

16. Archambeau C, Oppe M, Shen Y, Cornford D, Shawe-Taylor J. Variational inference for diffusion processes. *Adv Neural Inf Process Syst*. 2008;20:17-24.

17. Duan J. An Introduction to Stochastic Dynamics, Vol. 51. Cambridge University Press; 2015.

18. Cheng X, Wang H, Wang X, Duan J, Li X. Most probable transition pathways and maximal likely trajectories in a genetic regulatory system. *Phys A: Stat Mech Appl*. 2019;531:121779.

19. Ishikawa Y. *Stochastic Calculus of Variations*. de Gruyter; 2016.

How to cite this article: Dai M, Duan J, Hu J, Wang X. Nonparametric inference of stochastic differential equations based on the relative entropy rate. *Math Meth Appl Sci*. 2023;46(2):2986-2996. doi:10.1002/mma.8685

APPENDIX A: RELATIVE ENTROPY

The relative entropy or Kullback–Leibler divergence between the probability measures $P^g$ and $P^r$ of two solution processes for stochastic differential equation (1) with different drifts $g$ and $r$ is defined as
By the Girsanov theorem with $\theta = 0$, we obtain the Radon–Nykodim derivative

$$\frac{dQ}{dP} = \exp \left\{ \int_0^T u(Y_t) dB_t - \frac{1}{2} \int_0^T |u(Y_t)|^2 dt \right\}. \quad (A2)$$

Here, $\sigma(x)u(x) = g(x) - r(x)$, and the stochastic process $Y_t$ is governed by

$$dY_t = r(Y_t) dt + \sigma(Y_t) dB_t. \quad (A3)$$

Moreover, under the induced probability measure $Q$ by the Girsanov theorem, $Y_t$ also satisfies the stochastic differential equation

$$dY_t = g(Y_t) dt + \sigma(Y_t) dB_t^Q, \quad (A4)$$

where $B_t^Q = \int_0^t u(s) ds + B_t$ is a Brownian motion with respect to the induced probability measure $Q$. Then by weak uniqueness of the solutions for stochastic differential equation (1) and (A4), the likelihood function can be represented as

$$\ln \frac{P^g}{P^r} = \int_0^T u(Y_t) dB_t - \frac{1}{2} \int_0^T |u(Y_t)|^2 dt. \quad (A5)$$

To this end, we could use Itô formula to represent the stochastic integral as the Riemann integral and the relative entropy is

$$E_{P^r} \left[ \ln \frac{P^g}{P^r} \right] = \frac{1}{2} \int_0^T \int p^g(t, x) \left[ g(x) - r(x) \right] [D(x)]^{-1} \left[ g(x) - r(x) \right] dx dt, \quad (A6)$$

where $D = \sigma \sigma^T$ and $p^g(x, t)$ is the probability density of the stochastic differential equation (1) with drift vector field $g$. Assuming that $p^g(x, t)$ converges to the stationary density $p^g(x)$ for $t \to \infty$, we shall consider the relative entropy rate

$$d(P^g, P^r) = \lim_{T \to \infty} \frac{1}{T} E_{P^r} \left[ \ln \frac{P^g}{P^r} \right] = \frac{1}{2} \int p^g(x) \left[ g(x) - r(x) \right] [D(x)]^{-1} \left[ g(x) - r(x) \right] dx. \quad (A7)$$

Adding the Lagrange multiplier term, Equation (A7) becomes our optimal objective functional (11).