Classification of Stochastic Systems: Deep Learning and Hypothesis Testing

Qing Zhang, *Senior Member, IEEE*, Xiaohang Ma, and George Yin, *Life Fellow, IEEE*

***Abstract*—This article is devoted to classification of stochastic systems given by stochastic differential equations in continuous time. We develop two novel approaches. The first one is based on the use of a deep neural network (NN), whereas the second one uses hypothesis test-based methods. The idea of deep learning method focuses on treating the given stochastic system models by generating Monte Carlo sample paths. These samples are used to train a deep neutral network. A least square error is used as the loss function for network training. Then, the resulting weights are applied to out of sample Monte Carlo paths for testing. The underlying problem is then converted to a stochastic optimization task. Recursive stochastic algorithms are developed; convergence of the algorithm and rate of convergence are fully analyzed. Such deep NN approach compares favorably to the hypothesis test approaches. Then mean reversion models are studied to show the adaptiveness and power of our deep NN method. An advantage of the deep NN approach is real data can be used directly to train the deep NN. Therefore, model calibration can be bypassed all together.**

***Index Terms*—Deep neutral network (NN), hypothesis test, system classification.**

I. Introduction

THIS article develops novel identification methods for stochastic systems from a system classification point of view. The first method to be developed is built upon the use of deep neural networks (NNs). Using deep learning techniques, the essence of our approach is to convert the identification problem to a stochastic optimization procedure. Our second innovation is on the development of hypothesis test-based methods. It provides alternative computationally feasible procedures. We also develop the corresponding hypothesis tests for comparisons.

For both approaches, we begin with continuous-time systems given by stochastic differential equations. Our main task is to determine certain unknown parameters in the systems model based on observed data. As is well known, system identification have enjoyed numerous applications with much success from various fields such as engineering, physics, economics, biology. There has been a vast literature devoted to the subject. We cite the references [4], [7], [8], [13], [16], [21], [24], [26], [27], [33], [36], [37], [38], [39], [40]; see also related references for learning [34], [35] and statistical estimation [18] and references therein, which present a extensive research with a diverse and broad range of topics covered. For some of the recent study on identification of systems using least squares methods, we mention the work [28], [29], among others. In comparison with the existing literature on system identification, this article focuses on feasible computation schemes. While most of the existing work concentrates solely on an input–output approach with focus on treating discrete-time systems, we start with a continuous-time system given by a stochastic differential equation.

Initially, we delve into the application of deep NNs. Subsequently, we formulate hypothesis tests for the problem at hand. The essence of the deep learning methodology lies in initiating with predefined system models to generate Monte Carlo sample paths. These paths serve as the training data for a deep NN. The training process utilizes a nonlinear least square loss function. The resulting weights are then employed to analyze out-of-sample Monte Carlo paths. The performance of this deep NN approach proves to be superior to hypothesis test methods. We further investigate mean reversion models to showcase the adaptability and efficacy of our deep NN approach. A notable advantage of employing the deep NN method is its ability to directly utilize real data for training, thereby circumventing the need for model calibration entirely. The proposed algorithms have the potential of dealing with challenging system identification problems.

Deep Neutral Networks and Backpropagation: Neural networks are frequently employed to approximate complex and highly nonlinear functions that emerge from various applications. These networks are inherently compositional, relying on the composition of hidden layers containing base functions. A deep NN is commonly referred to as an NN with multiple hidden layers. In this article, one of our main focus is on deep NNs. For additional literature on deep NNs, we direct the reader to Nielsen's online book by Nielsen [31]. Some recent works on using stochastic gradient (SGD) method for deep learning, can be found in [1], [2], [6], [9]. [1], [2], [9] were mainly in the deterministic setting, whereas [6] was devoted to stochastic approximation algorithms with a constant stepsize. Estimation error depending on stepsize was derived with the use of L-mixing process.

Received 18 February 2025; accepted 10 April 2025. Date of publication 15 April 2025; date of current version 29 September 2025. This work was supported in part by the National Science Foundation under Grant DMS-2204240. Recommended by Associate Editor L. Bako. *(Corresponding author: George Yin.)*

Qing Zhang is with the Department of Mathematics, University of Georgia, Athens, GA 30602 USA (e-mail: qz@uga.edu).

Xiaohang Ma and George Yin are with the Department of Mathematics, University of Connecticut, Storrs, CT 06269 USA (e-mail: xiaohang.ma@uconn.edu; gyin@uconn.edu).

Digital Object Identifier 10.1109/TAC.2025.3561180

Backpropagation serves as a fundamental catalyst in the training of deep NNs. Broadly, Backpropagation is an algorithm employed for the supervised learning of artificial NNs through the utilization of gradient descent. Given an artificial NN and a loss (error) function, this method calculates the gradient of the loss function concerning the NN's weights.

The process involves backward propagation through the network, commencing with the computation of the gradient for the final layer of weights. To enhance computational efficiency, partial computations of the gradient from one layer are reused for the calculation of the gradient in the preceding layer. This backward flow of information is crafted to ensure the effective computation of the gradient at each layer.

Specifically, the backpropagation process necessitates three key components: (a) a data set comprising fixed pairs of input and output variables, (b) a feedforward NN with parameters specified by the weight w, and (c) a nonlinear loss (error) function, $L(w)$, defining the discrepancy between the desired output and the calculated output. In the context of this article, NN training employs the stochastic gradient descent method to determine the weight vector w that minimizes the loss function $L(w)$. Note that we use w in the above and in what follows. When we develop the algorithm, we use W_n to denote the realization (random noise involved).

Deep Learning-Based Method: This article centers around system identification using deep NNs. In the proposed model, our approach involves generating Monte Carlo samples, which are subsequently employed to train a deep NN. The state process serves as inputs to the NN, while the system index functions as the target.

The training process employs a nonlinear least square error between the target and the calculated output as the loss function, facilitating the determination of a weight vector for the network. Subsequently, this weight vector is applied to another set of Monte Carlo samples derived from an actual dynamic model. The resulting deep NN output provides an estimate for system identification.

In this article, we demonstrate the adaptiveness and effectiveness of our deep learning-based system identification through numerical examples. The deep learning-based approach compares favorably to the hypothesis tests in linear cases. Then, a mean reversion model is studied to show the flexibility of our deep learning approach. One advantage of the deep learning approach is real data can be used directly to train the deep NN. Therefore, model calibration can be bypassed all together in applications.

For related literature on system identification, we refer the reader to Ljung [24], Chen and Guo [7], and Chen and Zhao [8] among others; see also Gelman et al. [12] for a thorough exploration of Bayesian methods on parameter estimation; and Hayes [15] for related signal processing applications requiring system identification for tasks like noise reduction, filtering, and signal prediction.

System identification is closely related to the traditional parameter estimation and classification of dynamic systems. The basic idea is to use statistical methods to build mathematical models of dynamical systems from measured data. We refer the reader to Soderstrom and Stoica [33] for a comprehensive introduction to system identification techniques, including least squares methods and a maximum likelihood estimation approach in connection with system identification. Recently, Battistelli and Tesi [3] considered a classification problem for identifying dynamic systems based on their trajectories. They considered a model-based approach and a data-driven approach. Nevertheless, they studied only linear deterministic models and their focus was on observation matrices. In the literature, similar ideas were also used in pattern recognition and classification and related applications of a formal syntactic classification methods to the processing of pictorial data with the work of Fu and colleagues [11] as a representative. Our approach is to some extent related such classification methods, but on the other hand different from their setup. We emphasize the stochastic dynamic systems aspects, whereas [11] was concerned with pattern classification from a classical perspective. In connection with deep learning-based application in nonlinear filtering with adaptive learning rates, we refer the reader to a recent paper by Qian et al. [32].

Our main contributions of this article are as follows.

1) We develop a learning-based system identification method by using a deep NN; the basic idea is to generate Monte Carlo sample paths from the given stochastic dynamic systems. These samples are then used to train the NN. A least square error loss function is used for network training. Then, the resulting weights are feedforward to out of sample Monte Carlo paths for testing.
2) We also develop hypothesis test-based methods based on system trajectories. Both a fixed sample (quadratic variation) test and sequential (likelihood ratio) test are obtained. These tests provide benchmark comparisons of the deep learning-based approach with classical statistical approaches.
3) The deep learning task is converted to a stochastic optimization task. It consists of two steps. The first step uses the Euler–Maruyama procedure to start the approximation. Then using the computed results in Euler–Maruyama approximation, we develop recursive stochastic gradient algorithms; convergence and rate of convergence of the algorithm are fully analyzed.

The rest of this article is organized as follows. Section II begins with our deep learning-based approach. Section III develops the corresponding hypothesis tests. Numerical examples are provided in Section IV. Finally, Section V concludes this article.

II. Classification/Identification of Stochastic Systems

In contrast to the traditional system identification tasks where one typically begins with input and output sequences, we consider systems given by solutions of stochastic differential equations. Our primary concern is on developing numerical methods for the identification task. Naturally, we need to work with discrete-time systems. To facilitate the transition from continuous-time systems to that of discrete-time systems, we first use design a numerical procedure, to approximate the stochastic differential equations. With the sequences so obtained, we proceed with the development of the identification work.

To begin, let X_t be the state of a stochastic system satisfying

$$dX_t = b_\theta(X_t)dt + \sigma_\theta dB_t,\ X_0 = x \tag{1}$$

where θ is an unknown parameter taking values in $\{1, 2\}$, $\{b_i(x), i = 1, 2\}$ are functions of x, $\{\sigma_i,\ i = 1, 2\}$ are constants, and B_t is a standard Brownian motion. That is, we have two systems with different drift and diffusion coefficients in accordance with the value of θ, namely, System 1: (with $\theta = 1$) and System 2: (with $\theta = 2$). Given the observation of a segment of X_t, our goal is to determine which system it came from. More precisely, given $\{X_t :\ 0 \le t \le T\}$ for some T, one aims to determine the true value of θ.

In this article, we consider the observation in discrete time. Let $\eta > 0$ be the step size and $x_k = X_{k\eta}$. In the asymptotic analysis, we need to let $\eta \to 0$ and in the actual computation, it is a fixed positive constant. The x_k can be approximated by solution of the following difference equation:

$$x_{k+1} = x_k + \eta b_\theta(x_k) + \sqrt{\eta}\sigma_\theta v_k,\ x_0 = x \tag{2}$$

for $k = 0, 1, 2, \ldots, N_T$, with $N_T = \lfloor T/\eta \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x. Note that the above equation is simply an Euler–Maruyama algorithm of discrete-time approximation for (1). Here, $\{v_k\}$ is a sequence of i.i.d. (independent and identically distributed) Gaussian random variables with mean 0 and variance 1. In fact, we choose $\{v_k\}$ to be the Brownian increments, in that

$$v_k = \Delta B(\eta k) := \frac{B(\eta(k+1)) - B(\eta k)}{\sqrt{\eta}}.$$

A. Setup and Asymptotic Properties

Because we are dealing with recursive algorithms, a useful machinery to study the asymptotic properties is the method of weak convergence. In what follows, we first recall the notion of weak convergence and related issues. For simplicity, the discussion is kept in a relative short way. Further references and various terminology can be found in the reference of Kushner and Yin [23, ch. 7] and references therein.

Weak convergence: Let ψ_n and ψ be $\mathbb{R}^s$-valued random variables, we say ψ_n converges weakly to ψ, if and only if for any bounded and continuous function $f(\cdot)$, $\mathbb{E}f(\psi_n) \to \mathbb{E}f(\psi)$. In lieu of random variables in an Euclidean space, the definition can be extended to random elements in a metric space.

Tightness: ψ_n is said to be tight, if for each $\varepsilon > 0$, there is a compact set K_ε, such that $\mathbb{P}\{\psi_n \in K_\varepsilon\} \ge 1 - \varepsilon$, for all n. In fact, weak convergence and tightness can be extended to random variables taking values in some metric space. It is known that on a complete separable metric space (e.g., $\mathbb{R}^d$), the notion of tightness is equivalent to relative compactness. This is known as Prohorov's Theorem.

Skorohod representation: Let $\psi^\varepsilon(\cdot)$ converge to $\psi(\cdot)$ weakly in $D[0, \infty)$, which is the space of functions whose paths being right continuous and having left limits, equipped with appropriate weak topology (known as Skorohod topology). By a suitable choice of the probability space, the weak convergence becomes convergence w.p.1in the metric of $D[0, \infty)$. That is, there is a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, with $\{\tilde{\psi}^\varepsilon(\cdot)\}$, $\tilde{\psi}(\cdot)$ defined on it, such that for each Borel set A in $D[0, \infty)$

$$\tilde{\mathbb{P}}[\tilde{\psi}^\varepsilon(\cdot) \in A] = \mathbb{P}[\psi^\varepsilon(\cdot) \in A],\ \tilde{\mathbb{P}}[\tilde{\psi}(\cdot) \in A] = \mathbb{P}[\psi(\cdot) \in A]$$

and

$$\tilde{\psi}^\varepsilon(\cdot) \to \tilde{\psi}(\cdot) \text{ w.p.1}$$

as $\varepsilon \to 0$ in the topology of $D[0, \infty)$. In what follows, when we use this result, we often omit the use of the tilde notation for simplicity.

Euler–Maruyama algorithm: With the preparation above, we coming back to (2). Under suitable conditions, we can derive the convergence of $\{x_k\}$ together with the convergence rate. To this end, define a piecewise constant interpolation $x^\eta(t) = x_k$ for $t \in [k\eta, k\eta + \eta)$. Define also

$$v^\eta(t) = \sqrt{\eta} \sum_{k=0}^{\lfloor t/\eta \rfloor - 1} v_k.$$

We have the following result.

Lemma 2.1: Assume that for each $\theta \in \{1, 2\}$, $b_\theta(\cdot)$ is continuous and that system (1) has a unique solution (unique in the sense of in distribution). Then $x^\eta(\cdot)$ converges weakly to $X(\cdot)$ as $\eta \to 0$ such that $X(\cdot)$ is the solution of (1).

Remark 2.2: A proof of the above lemma can be found in [32, Prop. 1]. In fact, we dealt with nonlinear SDEs with a switching Markov chain there. To study the rate of convergence, we can show

$$\mathbb{E} \sup_{0 \le t \le T} |x^\eta(t) - X(t)|^2 = O(\eta^{1-\Delta}) \text{ for } 0 < \Delta < 1;$$

see [17]. To improve the rate, we can, in fact, examine the following continuous-time interpolation. Define

$$\overline{x}^\eta(t) = x^\eta(0) + \int_0^t b_\theta(x^\eta(s))ds + \sigma_\theta \int_0^t dw(s).$$

Then, we can show that

$$\mathbb{E} \sup_{0 \le t \le T} |\overline{x}^\eta(t) - X(t)|^2 = O(\eta); \tag{3}$$

see [25]. Thus, the Euler–Maruyama method provides us with a good discrete-time approximation of the continuous-time systems. Henceforth, we focus on the discrete time version of the system.

Deep Learning Approach. Setup: We propose our deep learning-based approach as follows. Let N_{seed} denote the number of training sample paths. We fix $\theta = 1, 2$. For a fixed $N \le N_T$ and a sample point ω, we take $\{x_1(\omega), x_2(\omega), \ldots, x_N(\omega)\}$ as the input vector to the NN and $\theta_0 = I_{\{\theta=1\}}$ (where I_A is the indicator of A) as the target, where $\theta_0 = 1$ if $\theta = 1$ and $\theta_0 = 0$ if $\theta = 2$.

Different from the usual deep learning approach for approximating a nonlinear function or finding the minimizers of an objective function, where one generates random samples, and then uses a stochastic gradient procedure to find the needed approximation, we have random data coming from the solutions of stochastic differential equations. Our main goal is the identification/classification of such stochastic systems use deep learning methods.

In our deep NN, the output is generated by processing input values through multiple layers. Each input is multiplied by its corresponding weight, taking linear combination with others, and adjusted by a bias. The result is then passed through an activation function. This process is repeated across each hidden layer, with the output of one layer serving as the input for the next, until the final output is produced. For technical details such as dependence of network output and loss function on input data, we refer the reader to Nielsen [31].

Let ξ denote the NN (final) output, which is real valued depending on a weight vector w and also affected by a vector-valued noise disturbance ζ. Here ζ represents a combined random effect resulted from numerical solutions of the SDE and other data processing. We assume that both w and ζ are valued in $\mathbb{R}^d$. The goal is to find the NN weights w to minimize a least square cost function

$$\overline{L}(w) = \frac{1}{2}\sum_{\theta=1}^{2}\mathbb{E}[\xi(w,\zeta)-\theta_0]^2 = \frac{1}{2}\sum_{\theta=1}^{2}\mathbb{E}[\xi(w,\zeta)-I_{\{\theta=1\}}]^2 \tag{4}$$

where ζ is the random disturbance appeared in the output with $\xi : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$. Note that the expected value in (4) cannot be calculated analytically. We only have noisy samples available. We thus compute the noisy sample gradient instead. Denote the noisy gradient estimate of $\overline{L}(w)$ by $\nabla L(w,\zeta)$, where $\zeta \in \mathbb{R}^d$ is the measurement noise appeared in the sample gradient calculation. Thus, $L : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ and

$$\nabla L(\xi(w,\zeta)) = \nabla_w L(\xi(w,\zeta)) = \sum_{\theta=1}^{2}\xi_w(w,\zeta)[\xi(w,\zeta)-\theta_0]$$

where $\nabla_w L(w,\zeta)$ denotes the gradient with respect to W and

$$\xi_w(w,\zeta) = \frac{\partial \xi(w,\zeta)}{\partial w} \in \mathbb{R}^d$$

denotes the gradient of ξ with respect to the weight vector W. Note that for simplicity, we assume the noisy gradient $\nabla L(w,\zeta)$ is available throughout this article, and comment on the possibility of using finite difference approximation at the end of this article in the concluding remarks. To find the "best match" for our parameter, we design a stochastic approximation type of algorithms. Following the back-propagation method in NNs to search the optimal weights, the stochastic approximation algorithm to be used throughout, takes the form

$$\begin{aligned} W_{n+1} &= W_n - \rho\nabla L(\xi(W_n,\zeta_n)) \\ &= W_n - \rho\sum_{\theta=1}^{2}\xi_w(W_n,\zeta_n)[\xi(W_n,\zeta_n)-\theta_0] \end{aligned} \tag{5}$$

where $\rho > 0$ is small enough serving as the learning rate.

In this article, the training dataset is divided into N_0 batches, i.e., the Monte Carlo data is divided into smaller and manageable subsets. Instead of processing the entire dataset at once, the model trains on these batches sequentially. In (5), a batch is taken at each time step when updating W_n. Following the training of the NN, the resulting weights are feedforward to out of sample Monte Carlo paths for testing.

Algorithm 1: Deep learning-based method.

```
Number of training samples: $N_{\text{seed}}$:
NN training input: $\{x_1(\omega),\ldots,x_N(\omega)\}$ with sample $\omega$
NN output: $\xi$
Loss function: $\overline{L}(w)$
NN training output: Weights $w$
Testing: Weight $w$ is used to feedforward on
  out-of-sample $\{x_1(\widetilde{\omega}),\ldots,x_N(\widetilde{\omega})\}$ leading to system
  mode estimate $\widetilde{\theta} = \widetilde{\theta}(\widetilde{\omega}) = 0, 1$.
```

Remark 2.3: Note that there is an alternative way of choosing the the sample loss function

$$L(\xi(w,\zeta)) = \frac{\sum_{\theta=1}^{2}\sum_{m=1}^{N_{\text{seed}}}|\xi(w,\zeta)-\theta_0|^2}{2N_{\text{seed}}}.$$

That is, we approximate the expectation using an sample average. However, it can be seen later that the stochastic approximation algorithm that we construct in fact does have an "averaging effect." So it is easier to construct an algorithm without the sample averaging.

To analyze the stochastic recursive algorithm, it is more efficient to examine its connection to continuous-time dynamic systems. Thus, we shall take a continuous-time interpolation $w^\rho(\cdot)$ that belongs to appropriate function space that is right continuous and have left limits with Skorohod type of topology. To gain further insight, we provide a road map below before proceeding to establish the convergence of the algorithm.

Road Map on Convergence and Rates of Convergence of Stochastic Gradient Algorithms: Our effort is to provide a solid foundation for our deep learning approach. Before diving into the technical details, we provide a road map, connect the dots, and indicate the way of technical development.

1) First, we prove the convergence of our deep learning-based numerical methods. This consists of the following multiple steps.
   a) To begin, we need to find numerical solutions of stochastic differential equations (SDEs). This is done by means of Euler–Maruyama methods. We mention the conditions needed and results to be obtained, but refer the technical details to our previous work in this part.
   b) Once the numerical solutions of the SDEs are obtained, we use them to construct the NN.
   c) The weights we construct of the NN is the parameter in our nonlinear loss function (nonlinear least square objective function).
   d) We need to obtain the optimizer of the objective function using stochastic gradient algorithms. So we construct such an algorithm.

e) To study the convergence of the stochastic gradient algorithm, we use the methods of martingale averaging through the associated deterministic ordinary differential equations. To obtain the convergence, because the iterates might not be bounded, we use a truncation device and first show that the truncated sequence has the desired convergence property. Then we show the untruncated sequence is also convergent. This is Theorem 2.6. The first step of the convergence proof establish the desired compactness (tightness).

f) The significance of the limit ordinary differential equation (ODE) is that the stationary point of it is precisely the minimizer of the objective function or the loss function.

2) Next, after the convergence of the numerical methods is established, we proceed to ascertain the convergence rates. This is done by study the centered and normalized error of the approximation sequence. We shall show that a normalized error sequence properly scaled convergence to a desired stochastic differential equations. The scaling factor together with the stationary covariance of the limit SDE gives us the rate of convergence. The results are in Theorem 2.11.

Remark 2.4: In the development above, we assume the uniqueness of the minimizer to simplify the discussion of the study. What if the minimizers are not unique. We provide a remark here. In view of the convergence and rates of convergence, one assumptions we used is that the minimizer of the nonlinear least square loss is unique. The convergence of the stochastic gradient algorithm will be obtained by showing a suitably interpolated sequence converges to a limit ODE, whereas the rate of convergence is carried out by examining a sequence of scaled and centered estimation errors. One question immediately follow. What if the minimizer is not unique. The iterates will converges not to a point, but to a set. Likewise, the rate of convergence can be studied using suitably scaled set valued sequences. In fact, the loss function need not be smooth. All the stated conditions presented can be relaxed. We refer this to our recent paper [30] for details. Note that in such cases, the ODEs need to be replaced by differential inclusions, and SDEs will need to be replaced stochastic differential inclusions. In fact, the stochastic differential inclusion is a new concept in the stochastic analysis literature. However, for the current article, we prefer to keep the current setting to present the main idea of the deep learning frame work.

Convergence of Recursive Algorithm: It is important to obtain the asymptotic properties of the recursive algorithm (5). To proceed, we define piecewise constant interpolation

$$W^{\rho}(t) = W_n \quad \text{for} \quad t \in [n\rho, n\rho + \rho).$$

To deal with the possible unboundedness, in lieu of (5), we work with a ν-truncation defined as follows. The idea is that if the iterate is within a ball of radius ν, we keep it as it is. If it is outside a ball of radius $\nu + 1$, we reset it as 0. If the iterates are in between the ball of radius ν and $\nu + 1$, we let it be a smooth function. To be more specific, for any fixed but otherwise arbitrary $\nu > 0$, define

$$W_{n+1}^{\nu} = W_n^{\nu} - \rho \sum_{\theta=1}^{2} \xi_w(W_n^{\nu}, \zeta_n)[\xi(W_n^{\nu}, \zeta_n) - \theta_0] q^{\nu}(W_n^{\nu}),$$
$$W_0^{\nu} = W_0 \tag{6}$$

where $q^{\nu}(w)$ is a truncation function that is smooth and that satisfies

$$q^{\nu}(w) = \begin{cases} 1 \text{ if } w \in O_{\nu} \\ 0 \text{ if } w \in \mathbb{R}^d - O_{\nu+1} \end{cases}$$

where $O_{\nu} = \{x \in \mathbb{R}^d : |x| \leq \nu\}$ is the ball in $\mathbb{R}^d$ with radius ν. To analyze the algorithm, we do not look at the discrete iterates directly. Rather we take a continuous-time interpolation and connect the iterates with continuous-time dynamic systems. We show that suitably interpolated sequence converges to a limit ODE. Then, the stationary point of the limit is precisely the optimal weight we are searching for. Define

$$W^{\rho}(t) = W_n, \; W^{\rho,\nu}(t) = W_n^{\nu} \quad \text{for} \quad t \in [n\rho, n\rho + \rho).$$

Then, $W^{\rho,\nu}(\cdot)$ is a ν-truncation of $W(\cdot)$ according to [23, p. 284], which means that $W^{\rho,\nu}(t) = W^{\rho}(t)$ for all t until the first exit time from the O_{ν}, the ball with radius ν. For the convergence of the proposed algorithm, we shall assume the following conditions. For the analysis to follow, we first establish that the truncated process is convergent. Then we will let $\nu \to \infty$ and show that the un-truncated process is also convergent.

(C1) The $\{\zeta_n\}$ is a bounded stationary uniform mixing sequence.

(C2) For each ζ, the function $\xi(\cdot, \zeta)$ together with its partial derivatives with respect to w up to the second order are continuous; for each ζ, $\xi(w, \zeta)$ has a polynomial growth in that $|\xi(w, \zeta)| \leq K_0(1 + |w|^p)$ for each w, each ζ, and some positive real number K_0, and some positive integer p.

(C3) The differential equation

$$\dot{w}(t) = -\sum_{\theta=1}^{2} \overline{\xi}_w(w(t))[\overline{\xi}(w(t)) - \theta_0] \tag{7}$$

has a unique (unique in distribution) solution for each initial condition $w(0) = W_0$.

Remark 2.5: Let us briefly discuss the conditions above. Condition (C1) is mild. For a definition of a uniform mixing sequence (which essentially uses a mixing measure ϕ_p with $p = \infty$), we refer the reader to [10, pp. 347–348]. The mixing condition essentially allows flexibility of correlated noise sequences, which is far more general than the usually assumed independent and identically distributed (i.i.d.) noise assumptions. More specifically, it requires the remote past and distant future being asymptotically independent. The correlation can involve infinitely many terms, but the correlation decays as the time increases.

Because $\xi(\cdot, \zeta)$ is a smooth function with respect to w, $\{\xi(w, \zeta_n) - \theta_0\}$ and $\{\xi_w(w, \zeta_n)\}$ are also bounded stationary sequence for each w. In view of the stationarity, and the smoothness of $\xi(\cdot,\zeta)$, denote

$$\mathbb{E}\xi(w,\zeta_n)=\overline{\xi}(w) \text{ and } \mathbb{E}\xi_w(w,\zeta_n)=\overline{\xi}_w(w).$$

A consequence of (C1) is $\mathbb{E}L(w,\zeta_n)=\overline{L}(w)$ for each $w\in\mathbb{R}^d$ with $\overline{L}(w)$ given by (4) and

$$\nabla\overline{L}(w)=\sum_{\theta=1}^{2}\overline{\xi}_w(w)[\overline{\xi}(w)-\theta_0]. \tag{8}$$

Another important consequence of the mixing condition is the mixing implies ergodicity [20, Remark 5.3, p. 488]. Thus, for each w and each positive integer m

$$\begin{aligned}&\frac{1}{n}\sum_{j=m}^{n+m-1}\xi_\ell(w,\zeta_j)\to\overline{\xi}(w) \text{ w.p.1 as } n\to\infty,\\&\frac{1}{n}\sum_{j=m}^{n+m-1}\xi_w(w,\zeta_j)\to\overline{\xi}_w(w) \text{ w.p.1 as } n\to\infty.\end{aligned} \tag{9}$$

In fact, the following slightly weaker conditions hold:

$$\begin{aligned}&\frac{1}{n}\sum_{j=m}^{n+m-1}\mathbb{E}_m\xi(w,\zeta_j)\to\overline{\xi}(w) \text{ in probability as } n\to\infty,\\&\frac{1}{n}\sum_{j=m}^{n+m-1}\mathbb{E}_m\xi_w(w,\zeta_j)\to\overline{\xi}_w(w) \text{ in probability as } n\to\infty\end{aligned} \tag{10}$$

where $\mathbb{E}_m$ denotes the conditional expectation conditioned on the information up to m (i.e., conditioned on the σ-algebra generated by $\{\xi(w,\zeta_k): k\le m\}$). Note that we only need a weak law of large number type condition as in (10) holds in our subsequent analysis for the convergence of the numerical algorithm.

Condition (C2) requires the sample cost $\xi(w,\zeta)$ grows polynomially with respect to w. It also poses some smoothness condition on the function and its partial derivatives. This condition then implies that $L(w,\zeta)$ has polynomial growth of order p. The growth condition posed here is rather general that includes the quadratic in w of $L(w,\zeta)$ and much more beyond that.

Taking into consideration of the mixing condition, the limit ordinary differential (7) is of the form

$$\dot{w}(t)=-\sum_{\theta=1}^{2}\overline{\xi}_w(w(t))[\overline{\xi}(w(t))-\theta_0]. \tag{11}$$

Condition (C3) requires the solution of the ODE having a unique solution. Sufficient condition ensuring (C3) can be readily derived, for example, if we require $\xi(\cdot,\zeta)$ and $\xi_w(\cdot,\zeta)$ to be locally Lipschitz in the first variable, then the uniqueness is verified. However, for our purpose of analysis, the uniqueness is only required to be in the sense of in probability distribution. Thus it is much weaker than the pathwise uniqueness.

Theorem 2.6: Under conditions (C1)–(C3), we have $W^{\rho,\nu}(\cdot)$ converges weakly to $w^\nu(\cdot)$ as $\rho\to 0$, where

$$\begin{aligned}&\dot{w}^\nu(t)=-\sum_{\theta=1}^{2}\overline{\xi}_w(w^\nu(t))[\overline{\xi}(w^\nu(t)-\theta_0]q^\nu(w^\nu(t)),\\&w^\nu(0)=W_0.\end{aligned} \tag{12}$$

Proof: The proof consists of two steps. In the first step, we show that the sequence $\{W^{\rho,\nu}(\cdot)\}$ is tight. Then in the second step, we characterize the limit by showing that $w^\nu(\cdot)$ is the solution of an appropriate martingale problem with a suitable operator.

Step 1: Tightness. Note that in view of the ν-truncation, the polynomial growth of $\xi(w,\zeta)$, W_n^ν and hence $W^{\rho,\nu}(\cdot)$ are bounded. Let $\mathbb{E}_t^{\rho,\nu}$ denote the conditional expectation conditioned on the σ-algebra generated by $\{W_k^\nu,\zeta_k: k\le\lfloor t/\rho\rfloor\}$, with $\lfloor t/\rho\rfloor$ denoting the integer part of t/ρ. In the calculation to follow, for notational simplicity, we often suppress the ν-dependence in W_k^ν and write it as W_k instead. We will retain the ν dependence notation when it is necessary. We will use this convention as long as there is no confusion. For any $t,s>0$

$$\begin{aligned}&\mathbb{E}_t^{\rho,\nu}|W^{\rho,\nu}(t+s)-W^{\rho,\nu}(t)|^2\\&=\mathbb{E}_t^{\rho,\nu}|W_{\lfloor(t+s)/\rho\rfloor}-W_{\lfloor t/\rho\rfloor}|^2\\&=\rho^2\mathbb{E}_{\lfloor t/\rho\rfloor}\sum_{k=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}\sum_{j=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}[\ell_{kj}+\ell'_{kj}]\end{aligned} \tag{13}$$

where

$$\begin{aligned}\ell_{kj}=&\left(\sum_{\theta=1}^{2}[\xi(W_j,\zeta_j)-\theta_0]q^\nu(W_j)\right)'\\&\times\sum_{\theta=1}^{2}[\xi(W_k,\zeta_k)-\theta_0]q^\nu(W_k)\end{aligned} \tag{14}$$

where ξ'_w denotes the transpose of ξ_w and ℓ'_{kj} denotes the transpose of ℓ_{kj}. Note that in the above, we have suppressed ν dependence for notational simplicity. It is easily seen that there is a random $\gamma_1^\rho(s)>0$ such that for some $K>0$

$$\begin{aligned}&\rho^2\mathbb{E}_{\lfloor t/\rho\rfloor}\sum_{k=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}\sum_{j=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}\ell_{kj}\\&\le K\rho^2\left(\frac{t+s}{\rho}-\frac{t}{\rho}\right)^2\\&\le\gamma_1^\rho(s).\end{aligned}$$

It then yields

$$\lim_{s\to 0}\limsup_{\rho\to 0}\mathbb{E}\gamma_1^\rho(s)=0. \tag{15}$$

Likewise

$$\rho^2\mathbb{E}_{\lfloor t/\rho\rfloor}\sum_{k=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}\sum_{j=\lfloor t/\rho\rfloor}^{\lfloor(t+s)/\rho\rfloor-1}\ell'_{kj}$$

$$\leq \gamma_2^\rho(s)$$

such that

$$\lim_{s\to 0}\limsup_{\rho\to 0}\mathbb{E}\gamma_2^\rho(s)=0. \tag{16}$$

Then (13), (14), (15), and (16) imply that the sequence $\{W^{\rho,\nu}(\cdot)\}$ is tight in $D[0,\infty)$ the space of functions defined on $[0,\infty)$ taking values $\mathbb{R}^d$ that are right continuous have left limits endowed with the Skorohod topology; see [22, Theorem 3, p.47].

Step 2: Characterization of the limit. Because $\{W^{\rho,\nu}(\cdot)\}$ is tight, it is sequentially compact. In accordance with the Prohorov's theorem [23, Chapter 7], we can extra a weakly convergent subsequence. Select such a sequence, and still denote the subsequence by $W^{\rho,\nu}(\cdot)$, and denote the limit by $w^\nu(\cdot)$. We shall show that $w^\nu(\cdot)$ is the solution of the martingale problem with operator $\mathcal{G}^\nu$. For any function f that has continuous partial derivatives with respect to w and that has a compact support, define the operator $\mathcal{G}$ as

$$\mathcal{G}^\nu f(w)=-\sum_{i=1}^{2} f'_w(w)\nabla\bar{\xi}_w(w)[\bar{\xi}(w)-\theta_0]q^\nu(w). \tag{17}$$

Since $W^{\rho,\nu}(\cdot)$ converges weakly to $w^\nu(\cdot)$, by the Skorohod representation [23, Chapter 7] with no change in notation, we may assume (with a slight abuse of notation), $W^{\rho,\nu}(\cdot)\to w^\nu(\cdot)$ w.p.1, and the convergence is uniform in any compact interval. We aim to show that

$$f(w^\nu(t))-f(w^\nu(0))-\int_0^t \mathcal{G}^\nu f(w^\nu(u))du \text{ is a martingale.}$$

To prove this, for any bounded and continuous function $h(\cdot)$, any positive integer m_0, any $t,s>0$, and $t_i\leq t$, we show

$$\begin{aligned}&\mathbb{E}h(w^\nu(t_i),i\leq m_0)\\&\times\left[f(w^\nu(t+s))-f(w^\nu(t))-\int_t^{t+s}\mathcal{G}^\nu f(w^\nu(u))du\right]=0.\end{aligned} \tag{18}$$

To verify (18), in turn, we start with the process $W^{\rho,\nu}(\cdot)$. By virtue of the weak convergence of $W^{\rho,\nu}(\cdot)$, the Skorohod representation, and the boundedness and continuity of $f(\cdot)$ and $h(\cdot)$, we have that as $\rho\to 0$

$$\begin{aligned}&\mathbb{E}h(W^{\rho,\nu}(t_i),i\leq m_0)[f(W^{\rho,\nu}(t+s))-f(W^{\rho,\nu}(t))]\\&\quad\to\mathbb{E}h(w^\nu(t_i),i\leq m_0)[f(w^\nu(t+s))-f(w^\nu(t))].\end{aligned} \tag{19}$$

For simplicity, write $\lfloor t/\rho\rfloor$ as t/ρ (likewise for $\lfloor (t+s)/\rho\rfloor$). Recall our convention of suppressing the ν-dependence and writing W_k^ν as W_k for simplicity. Choose n_ρ so that $n_\rho\to\infty$ as $\rho\to 0$ but $\delta_\rho=\rho n_\rho\to 0$. Partition $[t/\rho,(t+s)/\rho]$ using δ_ρ so that

$$\begin{aligned}&f(W^{\rho,\nu}(t+s))-f(W^{\rho,\nu}(t))\\&=\sum_{t\leq l\delta_\rho<t+s}[f(W_{ln_\rho+n_\rho})-f(W_{ln_\rho})]\\&=-\rho\sum_{\theta=1}^{2}\sum_{t\leq l\delta_\rho<t+s} f_w^{\nu,\prime}(W_{ln_\rho})\\&\quad\times\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}\nabla\xi_w(W_k,\zeta_k)[\xi(W_k,\zeta_k)-\theta_0]q^\nu(W_k)+o(1)\\&=-\rho\sum_{\theta=1}^{2}\sum_{t\leq l\delta_\rho<t+s} f_w^{\nu,\prime}(W^{\rho,\nu}(l\delta_\rho))\\&\quad\times\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}\nabla\xi_w(W^{\rho,\nu}l\delta_\rho),\zeta_k)[\xi(W^{\rho,\nu}(l\delta_\rho),\zeta_k)-\theta_0]\\&\quad\times q^\nu(W^{\rho,\nu}(l\delta_\rho))+o(1)\end{aligned} \tag{20}$$

where $o(1)\to 0$ in probability. To obtain the $o(1)$ term, we note that $f_w^\nu(\cdot)$, $\xi_w(\cdot,\zeta)$, and $[\xi(\cdot,\zeta)-\theta_0]$ are continuous functions. Letting $l\delta_\rho=\rho ln_\rho\to u$, then $\rho k\to u$ for any $ln_\rho\leq k<ln_\rho+n_\rho$. Inserting $\mathbb{E}_{ln_\rho}$ and using (10), we arrive at as $\rho\to 0$

$$\begin{aligned}&\mathbb{E}h(W^{\rho,\nu}(t_i),i\leq m_0)\\&\quad\times\mathbb{E}_{ln_\rho}\nabla\xi_w(W^{\rho,\nu}(l\delta_\rho),\zeta_k)[\xi(W^{\rho,\nu}(l\delta_\rho),\zeta_k)-\theta_0]\\&\quad\times q^\nu(W^{\rho,\nu}(l\delta_\rho))\\&\to\mathbb{E}h(w^\nu(t_i),i\leq m_0)\\&\quad\times\nabla\bar{\xi}_w(w^\nu(u))[\bar{\xi}(w^\nu(u))-\theta_0]q^\nu(w^\nu(u)).\end{aligned} \tag{21}$$

Then, (19) readily follows.

Corollary 2.7: Assume the conditions of Theorem 2.6 hold. Then, the untruncated sequence $\{W^\rho(\cdot)\}$ also convergence weakly. That is, $W^\rho(\cdot)$ converges weakly to $w(\cdot)$ such that $w(\cdot)$ is the unique solution of (7).

Idea of proof: We shall only comment on the proof briefly. Denote by $P^{w(0)}(\cdot)$ and $P^\nu(\cdot)$ the measures on the Borel subsets of $D[0,\infty)$ that are induced by $w(\cdot)$ and $w^\nu(\cdot)$, respectively. Then $P^{w(0)}(\cdot)$ is unique because of the uniqueness of the (7) hence the uniqueness of the associated martingale problem with operator $\mathcal{L}_o$. In addition, $P^{w(0)}(\cdot)$ coincide with $P^\nu(\cdot)$ on paths in $D[0,\infty)$ with values in O_ν for any $t\leq T$ for each $0<T<\infty$. Moreover, for any $0<T<\infty$

$$P^{w(0)}(\sup_{t\leq T}|w(t)|\leq\nu)\to 1 \text{ as } \nu\to\infty.$$

Putting the above together, by virtue of the weak convergence of $W^{\rho,\nu}(\cdot)$ to $w^\nu(\cdot)$, $W^\rho(\cdot)$ converges weakly to $w(\cdot)$. The uniqueness further implies that the chosen subsequence is irrelevant.

The above result is a limit for ρ small and n large but ρn remains bounded. Next, we consider the situation that $\rho n\to\infty$. Thus, it is a stability result.

Corollary 2.8: Suppose that there is a unique w^* that is an asymptotic stable point of (7), and than $\nabla\bar{\xi}_w(w^*)\nabla\bar{\xi}'_w(w^*)$ is invertible. In addition, $\{W_n\}$ is tight in $\mathbb{R}^d$. Then $W^\rho(\cdot+t_\rho)$ converges weakly to w^* as $\rho\to 0$, where $t_\rho\to\infty$ as $\rho\to 0$.

Remark 2.9: Note that in the above, we assumed the tightness of $\{W_n\}$. Sufficient conditions for the tightness can be provided similar to the second step of the proof of rate of convergence in the later section. We will simply omit the detail here and assume the tightness.

Proof: For arbitrary $0 < T < \infty$, consider a pair of sequences $\{W^\rho(\cdot + t_\rho), W^\rho(\cdot + t_\rho - T)\}$, which is tight as in the proof of Theorem 2.6. Select a weakly convergence subsequence and still index by ρ with limit $(w(\cdot), w_T(\cdot))$. Note that $w(0) = w_T(T)$. The precise value of $w_T(0)$ may be unknown, but all possible $W_T(0)$, for all T and for all convergent subsequences, belong to a tight set because $\{W_n\}$ is tight. This together with the stability of (7) and Theorem 2.6, for any $\eta > 0$ there exists a T_η sufficiently large such that for all $T > T_\eta$, there is an η-neighborhood N_η such that $\mathbb{P}(w_T(T) \in N_\eta) \geq 1 - \eta$. The desire results then follows.

Next, let us explore more on the fixed point w^*. Note that the invertibility of $\nabla\overline{\xi}_w(w^*)\nabla\overline{\xi}'_w(w^*)$ and

$$\sum_{\theta=1}^{2} \nabla\overline{\xi}_w(w^*)[\overline{\xi}(w^*) - \theta_0] = 0$$

lead to

$$\sum_{\theta=1}^{2} [\nabla\overline{\xi}_w(w^*)\nabla\overline{\xi}'_w(w^*)][\overline{\xi}(w^*) - \theta_0] = 0.$$

As a result, the stationary point w^* is a solution of

$$\sum_{\theta=1}^{2} [\overline{\xi}(w^*) - \theta_0] = 0$$

which by the assumption is unique.

Rate of Convergence of Recursive Algorithms: In this article, our convergence and rates of convergence include both Euler–Maruyama scheme and the stochastic gradient algorithm. This combined approach has not received the needed attention in the machine learning study. Now, the essence of studying convergence rate for the stochastic gradient algorithm is to examine the error term $W_n - w^*$. We study the rate of convergence in two steps. It reveals that this error term varies with respect to the stepsize of the order $\sqrt{\rho}$. In addition, because we are dealing with stochastic processes, not only do we need to examine the closeness of the W_n to w^*, but also we need to examine the variation of the estimation error $W_n - w^*$ dynamically (in a dynamic system point of view). In the first part, we carry out a local analysis that shows a suitable interpolation of $\{(W_n - w^*)/\sqrt{\rho}\}$ is a solution of a stochastic differential equation. The scaling factor $\sqrt{\rho}$ together with the asymptotic covariance of the diffusion will give us a rate of convergence result. In the first step, a main hypothesis posed is a moment bound. That is, for sufficiently large n, the sequence $\{(W_n - w^*)/\sqrt{\rho}\}$ is bounded in an appropriate sense. In the second step, we provide sufficient conditions leads to the moment bound, which uses an assumption: There is a Lyapunov function for (7) that is nonlinear but locally quadratic near w^*. We proceed with the analysis below.

Step 1: To continue, for notational simplicity, we denote $g(w,\zeta) = \sum_{\theta=1}^{2} \xi_w(w,\zeta)[\xi(w,\zeta) - \theta_0]$. Note that $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$. Using (5), we expand the recursion and write it as

$$\begin{aligned} W_{n+1} = W_n - \rho[g(w^*,\zeta_n) + g'_w(w^*,\zeta_n)(W_n - w^*)] \\ + O(\rho|g_{ww}(W_n^+,\zeta_n)||W_n - w^*|^2). \end{aligned} \quad (22)$$

Then, define $U_n = (W_n - w^*)/\sqrt{\rho}$, where W_n^+ is on the line segment joining W_n and w^*. We have

$$\begin{aligned} U_{n+1} = U_n - \sqrt{\rho}g(w^*,\zeta_n) - \rho g'_w(w^*,\zeta_n)U_n \\ + O(\rho^{3/2}|g_{ww}(W_n^+,\zeta_n)||U_n|^2). \end{aligned} \quad (23)$$

To carry out the analysis, it is again appropriate to use a truncation device as in the proof of Theorem 2.6. However, for purely notational simplicity and given we have already demonstrated how to use the truncation in Theorem 2.6, we will simply assume $\{U_n\}$ is bounded (even though it is not). Note that $\{\zeta_n\}$ is a bounded uniform mixing sequence by (C1), so $\{g(w^*,\zeta_n)\}$ is also a bounded uniform mixing sequence. We need the following additional conditions. After the conditions are stated, we state the rate of convergence result in Theorem 2.11.

(C4) Denote the mixing measure of $\{g(w^*,\zeta_n)\}$ by $\phi(k)$. We require $\sum_{k=1}^{\infty} \phi^{1/2}(k) < \infty$.

(C5) There is an m_ρ such that

$$\mathbb{E}|(W_n - w^*)/\sqrt{\rho}|^2 = O(1), \quad \text{for all } n \geq m_\rho. \quad (24)$$

With U_n defined above, taking into condition of (C5), we define piecewise constant interpolation

$$u^\rho(t) = U_n \ \text{ for } \ t \in [(n - m_\rho)\rho, (n - m_\rho)\rho + \rho). \quad (25)$$

We also define

$$B_n = -\sqrt{\rho} \sum_{k=m_\rho}^{n-1} g(w^*,\zeta_k) \ \text{ and}$$

$$B^\rho(t) = B_n \ \text{ on } \ t \in [(n - m_\rho)\rho, (n - m_\rho + 1)\rho). \quad (26)$$

We next state a lemma. This lemma is essentially a functional central limit theory for the underlying mixing process.

Lemma 2.10: Assume the conditions in Corollary 2.8, and (C4) and (C5). Then $B^\rho(\cdot)$ converges weakly to a Brownian motion $B(\cdot)$, whose covariance is given by Σt with

$$\begin{aligned} \Sigma = \mathbb{E}g(w^*,\zeta_0)g'(w^*,\zeta_0) + \sum_{k=1}^{\infty} \mathbb{E}g(w^*,\zeta_k)g'(w^*,\zeta_0) \\ + \sum_{k=1}^{\infty} \mathbb{E}g(w^*,\zeta_0)g'(w^*,\zeta_k). \end{aligned} \quad (27)$$

The lemma above is a result concerning the functional central limit of mixing sequence. A proof can be found in [10, Chapter 7] or [5]. We thus omit the verbatim argument.

To proceed, we remark that by the mixing condition of $\{\zeta_n\}$, and noting $\mathbb{E}g_w(w^*,\zeta_n) = \overline{g}_w(w^*)$, for any positive integer m, we have

$$\frac{1}{n} \sum_{k=m}^{m+n-1} \mathbb{E}_m g_w(w^*,\zeta_k) \to \overline{g}_w(w^*) \text{ in probability as } n \to \infty \quad (28)$$

where $\mathbb{E}_m$ denotes the conditional mean with respect to the σ-algebra of the past information up to m (the σ-algebra generated by $\{W_0, \zeta_j : j \leq m\}$).

Theorem 2.11: Assume the conditions in Corollary 2.8, and (C4) and (C5). then $u^\rho(\cdot)$ converges weakly to $u(\cdot)$ so that $u(\cdot)$ is the solution of

$$du = -\overline{g}'_w(w^*)udt + dB(t). \tag{29}$$

Remark 2.12: Because the stochastic differential equation (29) is linear in u, it has a unique solution for each initial condition. In fact, for our convergence analysis, we only need the SDE (29) having a unique solution in the sense of in distribution.

Idea of proof of Theorem 2.11: Note that (29) may also be written as

$$du = -\overline{g}'_w(w^*)udt + \Sigma^{1/2}d\widetilde{B}(t) \tag{30}$$

where $\widetilde{B}(t)$ is a standard Brownian motion. The approach we use is martingale averaging. The associated operator for (30) is

$$\mathcal{L}H(w) = -\nabla H'(w)\overline{g}'_w(w^*)u + \frac{1}{2}\text{tr}(\Sigma\nabla^2 H(w))$$

for any H whose partial derivatives with respect to w up to the second order are continuous. We note that in the last term of (23), because $\{\zeta_n\}$ is a bounded uniform mixing sequence, and W_n^+ is close to w^*, the continuity of g together with its partial derivatives up to the second order implies the term $g_{ww}(W_n^+, \zeta_n)$ is bounded. Thus

$$\mathbb{E}\rho^{3/2}|g_{ww}(W_n^+, \zeta_n)||U_n|^2 \le K\rho^{3/2}\mathbb{E}|U_n|^2 = O(\rho^{3/2}).$$

Then

$$\sum_{k=t/\rho}^{(t+s)/\rho} K\rho^{3/2} = O(\rho^{1/2}) \to 0 \text{ as } \rho \to 0.$$

Thus, the last term in (23) does not contribute anything to the limit. So we need only examine the first line in (23). Using Lemma 2.10, we have the convergence to the Brownian motion $\widetilde{B}(\cdot)$ with covariance Σ. We can show use similar idea as in the proof of Theorem 2.6, the other term contributes to the limit drift term as desired. A few details are omitted.

Step 2: Verify $\mathbb{E}|U_n|^2 = O(1)$ for sufficiently large n. Here, we provide sufficient conditions that ensure the 2nd moment bound of U_n. It uses a Lyapunov function based approach. We shall use the following conditions.

(C6) Denote $\widetilde{w} = (w - w^*)$. There is a Lyapunov function $V : \mathbb{R}^d \mapsto \mathbb{R}$ satisfying $V(\widetilde{w})$ is locally quadratic in that $V(\widetilde{w}) = \widetilde{w}'\Gamma\widetilde{w} + o(|\widetilde{w}|^2)$ for some symmetric and positive definite matrix $\Gamma \in \mathbb{R}^{d\times d}$ such that $V(\widetilde{w}) \ge 0$ for all $\widetilde{w}$, $|V_w(\widetilde{w})| \le K(1 + V(\widetilde{w}))$, $V_{ww}(\cdot)$ is bounded, and that there is a $\lambda > 0$ such that

$$-V'_w(\widetilde{w})g_w(w^*)\widetilde{w} \le -\lambda V(\widetilde{w}).$$

As in Step 1, we can show that $g_{ww}(W_n^+, \zeta_n)$ is bounded. Direct calculation reveals that

$$\begin{aligned}
&\mathbb{E}_n V(\widetilde{W}_{n+1}) - V(\widetilde{W}_n)\\
&\quad = -\rho V'_w(\widetilde{W}_n)g(w^*, \zeta_n) - \rho V'_w(\widetilde{W}_n)g'_w(w^*)\widetilde{W}_n\\
&\qquad + \rho\mathbb{E}_n O(V'_w(\widetilde{W}_n)|g_{ww}(W_n^+, \zeta_n)||\widetilde{W}_n|^2)\\
&\qquad + \rho\mathbb{E}_n O(V_{ww}(\widetilde{W}_n^{++})|\widetilde{W}_{n+1} - \widetilde{W}_n|^2)\\
&\quad \le -\rho V'_w(\widetilde{W}_n)g(w^*, \zeta_n) - \rho\lambda V(\widetilde{W}_n)\\
&\qquad + o(\rho)(1 + V(\widetilde{W}_n))
\end{aligned} \tag{31}$$

where $\widetilde{W}_n^{++}$ is on the line segment between W_n and W^*. Define a perturbation of the Lyapunov function as

$$V_1(\widetilde{W}_n, n) = -\rho\sum_{k=n}^{\infty}\mathbb{E}_n V'_w(\widetilde{W}_n)g(w^*, \zeta_k).$$

The purpose of the perturbation is that it is small compared to the Lyapunov function V. In addition, it will result in the needed cancellation. We will demonstrate these below. It can be seen that

$$\begin{aligned}
|V_1(\widetilde{W}_n, n)| &\le \rho|V'_w(\widetilde{W}_n)||\sum_{k=n}^{\infty}\mathbb{E}_n g(w^*, \zeta_k)|\\
&\le O(\rho)(1 + V(\widetilde{W}_n)).
\end{aligned} \tag{32}$$

That is, the perturbation is small. We next show that it results in the right cancellation. In fact, we have

$$\begin{aligned}
&\mathbb{E}_n V_1(\widetilde{W}_{n+1}, n+1) - V_1(\widetilde{W}_n, n)\\
&\quad = \mathbb{E}_n V_1(\widetilde{W}_{n+1}, n+1) - \mathbb{E}_n V_1(\widetilde{W}_n, n+1)\\
&\qquad + \mathbb{E}_n V_1(\widetilde{W}_n, n+1) - V_1(\widetilde{W}_n, n)\\
&\quad = \rho V'_w(\widetilde{W}_n)g(w^*, \zeta_k)\\
&\qquad + O(\rho^2)(1 + V(\widetilde{W}_n)).
\end{aligned} \tag{33}$$

Define

$$\widetilde{V}(\widetilde{W}, n) = V(\widetilde{W}_n) + V_1(\widetilde{W}_n, n).$$

We obtain $\rho|V_w(\widetilde{W}_n)||\widetilde{W}_n|^2 \le o(\rho)V(\widetilde{W}_n)$. Then, it follows:

$$\begin{aligned}
&\mathbb{E}_n\widetilde{V}(\widetilde{W}_{n+1}, n+1) - \widetilde{V}(\widetilde{W}_n, n)\\
&\quad \le -\lambda V(\widetilde{W}_n) + O(\rho^2)(1 + V(\widetilde{W}_n))\\
&\qquad + o(\rho)V(\widetilde{W}_n).
\end{aligned} \tag{34}$$

Upper bound $O(\rho^2)(1 + V(\widetilde{W}_n))$ by $O(\rho^2)(1 + \widetilde{V}(\widetilde{W}_n, n))$, and likewise bound $o(\rho)V(\widetilde{W}_n)$ by $o(\rho)\widetilde{V}(\widetilde{W}_n, n)$. It can shown that

$$\mathbb{E}_n\widetilde{V}(\widetilde{W}_{n+1}, n+1) \le K\left(1 - \frac{\lambda}{2}\right)\widetilde{V}(\widetilde{W}_n, n) + O(\rho^2). \tag{35}$$

Iterating on the resulting estimate above and taking expectation, we arrive at

$$\begin{aligned}
&\mathbb{E}\widetilde{V}(\widetilde{W}_{n+1}, n+1)\\
&\le K\left(1 - \frac{\lambda}{2}\right)^n\mathbb{E}\widetilde{V}(\widetilde{W}_0, 0) + K\sum_{k=0}^{n}\left(1 - \frac{\lambda}{2}\right)^{n-k}O(\rho^2)\\
&\le K\left(1 - \frac{\lambda}{2}\right)^n\mathbb{E}\widetilde{V}(\widetilde{W}_0, 0) + O(\rho).
\end{aligned} \tag{36}$$

Choose m_ρ to be a positive integer so that $K\exp(-(\rho\lambda/2)m_\rho) < -K\rho$ and hence for all $n \geq m_\rho$

$$K\exp(-(\rho\lambda/2)n) \leq K\exp(-(\rho\lambda/2)m_\rho) < -K\rho.$$

Using (32), we further obtain

$$\mathbb{E}V(\widetilde{W}_n) \leq K\exp(-(\rho\lambda/2)n) + K\rho = O(\rho) \tag{37}$$

as desired. We used K as a generic positive constant with the notation $K > 0$, $K + K = K$ and $KK = K$ throughout the article. Up to now, Theorem 2.11 is proved.

Nonunique Stable Points of (7): To proceed, we generalize the result of Corollary 2.8. The generalization will be done in two ways. First, we assume the smoothness of the dynamics, but we replace the unique stable point of (7) by a finite number of isolated stable points. Then, we consider an even more general case, where no smoothness or continuity is assumed. The limit differential equation is replaced by a differential inclusion. Here, the usual SGD is replaced by SSD (stochastic subgradient descent).

Multiple Isolated Stable Points: Proposition 2.13: Assume the conditions of Corollary 2.8, but replace the existence of unique stable point w^* by that there are isolated stable points $w^{*,i}$ and stable matrices $-G^i$ for $i = 1, \ldots, l_0$ satisfying

$$\begin{aligned}&-\sum_{\theta=1}^{2}\overline{\xi}_w(w)[\overline{\xi}(w) - \theta_0]\\ &= -\sum_{i=1}^{l_0}[G^i(w - w^{*,i}) + o(|w - w^{*,i}|^2)].\end{aligned} \tag{38}$$

Then, $W^\rho(\cdot + t_\rho)$ converges weakly to $\widetilde{S} = \{w^{*,1}, \ldots, w^{*,l_0}\}$ in the sense $\mathrm{d}(w, \widetilde{S}) = \inf_{w^{*,j}\in\widetilde{S}}|w - w^{*,j}| \to 0$.

Idea of proof: We will be brief here. As in the proof of Corollary 2.8, for arbitrary $0 < T < \infty$, we still consider $\{W^\rho(\cdot + t_\rho), W^\rho(\cdot + t_\rho - T)\}$. We can use a truncation device as before, but for simplicity and without loss of generality, we simply suppress the ν notation and assume that the sequence is bounded. Then, the pair of sequences is tight with limit denoted by $(w(\cdot), w_T(\cdot))$. Suppose $W^\rho(\cdot + t_\rho)$ does not converges to $\widetilde{S}$. Then, there exists a subsequence $\{W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}})\}$ satisfying $\mathrm{d}(W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}}), \widetilde{S}) \not\to 0$ in probability, so $W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}}) \to \overline{w}$ for some $\bar{w} \notin \widetilde{S}$. The limit of $\{W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}}), W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}} - T)\}$ is still $(w(\cdot)), w_T(\cdot))$. Denote by $\widetilde{\mu}_i$ the minimal eigenvalues of G^i for $i = 1, \ldots, l_0$. Due to the conditions of this proposition, without loss of generality, assume G^1 is the matrix so that $\widetilde{\mu}_1 > \widetilde{\mu}_j$ for $j = 2, \ldots, l_0$. Note that we only have a finite number of isolated stable points. Thus, we obtain

$$\begin{aligned}w_T(T) = w(0) &= e^{-G^1T}w_T(0)\\ &+ \sum_{i=1}^{l_0}\int_0^T e^{-G^1(T-t)}[-G^i(w_T(t) - w^{*,i})\\ &+ o(|w_T(t) - w^{*,i}|^2)]dt.\end{aligned} \tag{39}$$

Furthermore, we obtain

$$\begin{aligned}w_T(T) - w^{*,1} &= e^{-G^1(T)}w_T(0) + o_T(1)\\ &+ \int_0^T e^{-G^1(T-t)}\left[-\left[\sum_{i=2}^{l_0}G^i(w_T(t) - w^{*,i})\right.\right.\\ &\left.\left. + \ o(|w_T(t) - w^{*,i}|^2)\right]\right] dt\end{aligned} \tag{40}$$

where $o_T(1) \to 0$ in probability as $T \to \infty$. Taking vector norm on both sides, using Gronwall's inequality, and for sufficiently large T, we can make $|w_T(T) - w^{*,1}|$ sufficiently small. This yields a contradiction to $\mathrm{d}(W^{\widetilde{\rho}}(\cdot + t_{\widetilde{\rho}}), \widetilde{S}) \not\to 0$ in probability.

Stochastic Subgradient Descent (SSD): Next, we further relax the condition on the smoothness of $\xi(\cdot, \zeta)$. In lieu of the usual SGD, we deal with a class of stochastic subgradient descent (SSD) algorithms. Recall that for a real-valued function $H(\cdot)$ defined on $\mathbb{R}^d$, a vector γ is a subgradient of $H(\cdot)$ at x if $H(x + y) - H(x) \geq \gamma' y$ for all $y \in \mathbb{R}^d$. We rewrite (5) as set-valued stochastic iterates. By upper semicontinuity of a set-valued function $H(x)$, roughly, we mean that $\lim_{y\to x} H(y) \subset H(x)$. More precisely, denoting by $N_\delta(x)$ a δ neighborhood of x, the upper semicontinuity $H(x)$ requires that $\cap_{\delta>0}\mathrm{co}[\cup_{y\in N_\delta(x)}H(y)] = H(x)$, where $\mathrm{co}(A)$ denotes the closed and convex hull of A. Now, rewrite (5) as

$$W_{n+1} = W_n - \rho\gamma_n(W_n, \zeta_n) \tag{41}$$

where $\{\gamma_n(\cdot, \cdot)\}$ is a sequence of set-valued mappings representing subgradient estimates. Denote by $\mathsf{SG}(x)$ the set of subgradient (also called the set of subdifferentials at x) of $H(\cdot)$ at x. The set is closed and convex and satisfies $\mathsf{SG}(x) = \cap_{\delta>0}\mathrm{co}[\cap_{y\in N_\delta}(x)\mathsf{SG}(y)]$. It is thus upper semicontinuous.

Proposition 2.14: Consider (41). In lieu of (C2) and (C3), assume that (a) $-\gamma_n \in G(W_n) + z(W_n, \zeta_n) + \widetilde{\psi}_n$, where $G(w) \subset O_{r_0}$ for each $w \in \mathbb{R}^d$ with O_{r_0} denoting a ball of radius $0 < r_0 < \infty$; (b) $\{\zeta_n\}$ satisfies (C1) so that $\mathbb{E}z(w, \zeta_n) = 0$; (c) $z(\cdot, \zeta) \in \mathbb{R}^d$ satisfies that for each bounded random variable W, $\lim_{\rho,\delta,n}\frac{1}{n_\rho}\sum_{k=n}^{n+n_\rho-1}\mathbb{E}\sup_{|Y|\leq\delta}|z(W + Y, \zeta_k) - z(W, \zeta_k)| = 0$; (d) $\{\widetilde{\psi}_n\}$ is another stationary sequence of $\mathbb{R}^d$-valued mixing random variables with mixing measure for some $p > 2$ defined by $\phi_p(\mathcal{G}|\mathcal{H}) = \sup_{A\in\mathcal{H}}|\mathbb{P}(A|\mathcal{G}) - \mathbb{P}(A)|$ ([10, pp. 345–346]) such that $\mathbb{E}|\widetilde{\psi}_n|^p < \infty$, $\mathbb{E}\widetilde{\psi}_n = 0$. Define $W^\rho(\cdot)$ as $W^\rho(t) = W_n$ for $t \in [\rho n, \rho n + \rho)$. Then, $W^\rho(\cdot)$ converges weakly to $w(\cdot)$ that is a solution of the following differential inclusion

$$\dot{w}(t) \in G(w(t)). \tag{42}$$

Idea of proof: We can still use the truncation techniques, but for notational simplicity, we suppress the ν and simply assume the iterates are bounded. The tightness of $\{W^\rho(\cdot)\}$ can be proved similarly as before. We can then extract convergent subsequence still denoted by $\{W^\rho(\cdot)\}$ for notational simplicity. Denote limit by $w(\cdot)$. By virtue of (c), we can replace $\frac{1}{n_\rho}\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}z(W_k, \zeta_k)$ by $\frac{1}{n_\rho}\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}z(W_{ln_\rho}, \zeta_k)$ with an error terms goes to zero in probability. Similar to the proof of Theorem 2.6, (more specifically similar to) (20) (using the same

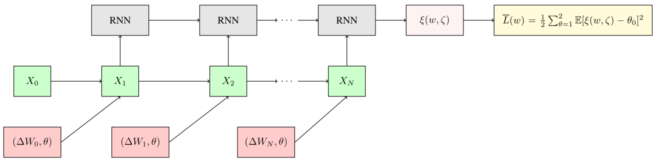


Fig. 1. RNN diagram.

convention), we need only consider

$$\frac{1}{n_\rho}\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}\mathbb{E}_{ln_\rho}z(W_{ln_\rho},\zeta_k) \text{ and } \frac{1}{n_\rho}\sum_{k=ln_\rho}^{ln_\rho+n_\rho-1}\mathbb{E}_{ln_\rho}\widetilde{\psi}_k. \tag{43}$$

Both the nonadditive and additive noise terms above can be averaged out to 0 due to the ergodicity (because of the mixing conditions). With a few details omitted, we obtain (42) in the limit.

Remark 2.15: Note that there are two mixing sequences in Proposition 2.14. The sequence $\{\zeta_n\}$ is a bounded uniformly mixing sequence so $z(w,\zeta_n)$ is a nonadditive uniform mixing noise, whereas $\{\widetilde{\psi}_n\}$ is a sequence of additive unbounded mixing sequence with pth moment. Note that vector-valued sequences in (43) can be replaced by set-valued sequence, then the limit will be in the sense with the use of distance function. Rather, we prefer to use simpler notation here. In fact, we used the usual convention that for a set A and $v \in \mathbb{R}^d$, $A+v$ is interpreted as $A+v=\{x+v : \text{for every } x \in A\}$. The proof of the above result uses weak convergence and averaging but replace the $\mathbb{R}^d$-valued random variables by set-valued random variables. Suppose that there is a locally asymptotically stable set $\widehat{S}$ (in the sense of Lyapunov) for the solutions of the differential inclusion (42). Then, it can be demonstrated that $\mathrm{d}(W^\rho(\cdot+t_\rho),\widehat{S}) \to 0$ in probability as $\rho \to 0$. Furthermore, we can examine a scaled sequence of the estimation error and obtain a stochastic differential inclusion limit; see [30, Theorem 3.1]. Regarding Proposition 2.13, it is possible to study the distribution of the multistable points; [19], in which one uses i.i.d. assumption together with the distribution of the noise. However, this is not our interest here.

Continuing on the development of learning: We have demonstrated that the algorithm that designed lead to the desired limit. Following the training of the NN and our asymptotic study, we have generate an optimal weight W. Then, this weight matrix will be applied to feedforward on out-of-sample data point $\widetilde{\omega}$ with the actual observation $\{x_n(\widetilde{\omega})\}$ as inputs in the subsequent testing stage which leads to NN output $\widetilde{\theta}_0(\widetilde{\omega})$. Define $\widetilde{\theta}=1$ if $\widetilde{\theta}_0 > 1/2$ and $\widetilde{\theta}=2$ if $\widetilde{\theta}_0 \le 1/2$. This $\widetilde{\theta}$ gives a deep NN-based estimator for θ.

Example 2.16: We use random seed $\eta_{\text{seed}} = 1 \times 10^5$ for initialization of the parameters in our Recurrent Neural Network (RNN) model; see Fig. 1. We trained the RNN using 10 000 samples with batch size 1024 for 150 epochs. The network has six layers, including one input layer, four hidden layers, and one output layer. The input layer designed to match the shape of the training data, following the input layer, there are four Simple RNN layers with 64, 32, 32, and 16 units (neurons), respectively. The network concludes with a dense output layer consisting of 2 units and a softmax activation function for our classification task. Also, for all hidden layers, we use the sigmoid activation function

$$\phi(x) = \sigma(x) = \frac{1}{1+e^{-x}}.$$

Next, we take $b_1(x) = b_2(x) = x$, and $\eta = 0.01$ and vary N and σ_i to examine the accuracy of our NN estimator. First, we fix $\sigma_2 = 0.5$ and vary σ_1. Then, we fix $\sigma_1 = 0.3$ and vary σ_2. The results are given in Tables I and II , respectively.

TABLE I
DEPENDENCE OF ACCURACY ON N AND σ_1 WITH FIXED $\sigma_2 = 0.5$

N \ σ_1	0.1	0.2	0.3	0.4	0.5
10	99.10%	96.07%	84.07%	67.20%	50.00%
20	99.50%	98.37%	92.97%	72.93%	50.00%
30	99.73%	99.60%	96.60%	77.70%	50.00%
40	99.83%	99.03%	97.70%	82.50%	50.00%
50	99.87%	99.57%	98.70%	84.17%	50.00%
60	99.93%	99.83%	98.50%	87.37%	50.00%
70	99.73%	99.67%	98.50%	87.57%	50.00%
80	99.83%	99.63%	98.63%	90.17%	50.00%
90	99.80%	99.87%	98.93%	91.10%	50.00%
100	99.87%	99.70%	99.17%	89.87%	50.00%

TABLE II
DEPENDENCE OF ACCURACY ON N AND σ_2 WITH FIXED $\sigma_1 = 0.3$

N \ σ_2	0.3	0.4	0.5	0.6	0.7
10	50.00%	70.27%	85.30%	92.30%	94.50%
20	50.00%	78.80%	93.00%	97.07%	98.53%
30	50.00%	84.10%	96.07%	98.70%	98.83%
40	50.00%	86.90%	97.87%	99.00%	99.30%
50	50.00%	90.20%	98.57%	98.90%	99.53%
60	50.00%	91.97%	98.73%	99.40%	99.53%
70	50.00%	90.77%	98.53%	99.20%	99.33%
80	50.00%	95.07%	99.00%	99.67%	99.73%
90	50.00%	93.77%	98.70%	99.10%	99.70%
100	50.00%	95.73%	99.27%	99.73%	98.97%

It can be seen from Tables I and II that the estimation accuracy improves in N. In addition, the farther apart the σ_1 and σ_2, the better the algorithm performs. It becomes difficult to separate Systems 1 and 2 when the σ's are moving together.

Note that the training stage is the most time consuming part. Normally, it takes a few thousands samples to train the network. The good part is that such computationally heavy stage is done offline. The feedforward part is simple and fast.

III. HYPOTHESIS TESTS

In this section, we design the corresponding hypothesis tests and compare their performance. We impose the following conditions.

(H1) For $i=1,2$, $b_i(x) = b_i x + O(1)$, for constants b_i, where $|O(1)| \le C$ for some constant C.

(H2) $\sigma_1 \ne \sigma_2$.

Without loss of generality, we assume $\sigma_1^2 < \sigma_2^2$.

Fixed sample size (quadratic variation) test: Let N denote the sample size. We consider the test statistics

$$Z_N = \frac{1}{\eta N}\sum_{k=1}^{N}(x_{k+1}-x_k)^2.$$

Then, we have

$$\begin{aligned} Z_N &= \frac{1}{\eta N}\sum_{k=1}^{N}(\eta b_\theta(x_k)+\sqrt{\eta}\sigma_\theta w_k)^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}(\sqrt{\eta}b_\theta(x_k)+\sigma_\theta w_k)^2. \end{aligned}$$

We use $\mathcal{H}_i$ to denote the hypothesis $\{\theta = i\}$ for $i = 1, 2$.

Description of the test: We accept $\mathcal{H}_1$ if

$$\left|\frac{Z_N}{\sigma_1^2}-1\right| \leq \left|\frac{Z_N}{\sigma_2^2}-1\right|$$

and accept $\mathcal{H}_2$ otherwise.

Let

$$\Gamma = 2\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2).$$

Then, the above inequality is equivalent to $Z_N \leq \Gamma$.

Proposition 3.1: Under Assumptions (H1) and (H2), there exist k_0 and $\eta_0 > 0$ such that the error probabilities

$$\begin{aligned} &P(Z_N \geq \Gamma|\mathcal{H}_1) \leq \eta \text{ and} \\ &P(Z_N \leq \Gamma|\mathcal{H}_2) \leq \eta \end{aligned}$$

for $N \geq k_0|\log\eta|$ and $0 < \eta < \eta_0$.

Proof: We only prove the first inequality. The proof for the second one is similar. Assume $\mathcal{H}_1$ holds. Let $a_k = \sqrt{\eta}b_1^2(x_k) + 2b_1(x_k)\sigma_1 w_k$. Then

$$Z_N = \frac{1}{N}\sum_{k=1}^{N}\left(\sqrt{\eta}a_k + \sigma_1^2 w_k^2\right).$$

Moreover, under (H1), there exists a constant C_0 such that

$$|a_k| \leq C_0(1+x_k^2+w_k^2), \quad \text{for} \quad \eta \leq 1.$$

Note that

$$\begin{aligned} &P(Z_N \geq \Gamma|\mathcal{H}_1) \\ &= P(Z_N - \Gamma \geq 0|\mathcal{H}_1) \\ &= P(Z_N - \sigma_1^2 \geq \alpha|\mathcal{H}_1), \text{ with } \alpha = \Gamma - \sigma_1^2 > 0 \\ &\leq P\left(\frac{\sqrt{\eta}}{N}\sum_{k=1}^{N}a_k \geq \frac{\alpha}{2}|\mathcal{H}_1\right) + P\left(\frac{\sigma_1^2}{N}\sum_{k=1}^{N}(w_k^2-1) \geq \frac{\alpha}{2}\right) \\ &:= I_1 + I_2. \end{aligned}$$

Using [14, Lemma 2.1], we can show $Ea_k^4 \leq C_1$ for some C_1. The Markov's inequality yields

$$I_1 \leq \frac{1}{(\alpha/2)^4}\frac{1}{N}\sum_{k=1}^{N}E\alpha_k^4 \leq C_1\eta^2.$$

Finally, standard large deviation estimate leads to, for some $C_2 > 0$

$$I_2 \leq C_2 e^{-C_2 N} \leq C_2 e^{C_2 k_0 \log\eta} = C_2\eta^2$$

with $k_0 = 2/C_2$. Take

$$\eta_0 = \min\{1, 1/(2C_1), 1/(2C_2)\}$$

to conclude the proof. □

Note that the condition $N \geq k_0|\log\eta|$ is required so that the error probabilities are less than η. More elaborated estimates for N^* are provided in the numerical section.

Sequential (likelihood ratio) test: We take the following test statistics:

$$S_n = \log\left[\frac{\Pi_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_{k+1}-x_k)^2}{2\eta\sigma_1^2}\right)}{\Pi_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_{k+1}-x_k)^2}{2\eta\sigma_2^2}\right)}\right].$$

Let

$$c_1 = \log(\sigma_2/\sigma_1) \text{ and } c_2 = (1/\sigma_2^2 - 1/\sigma_1^2)/2.$$

Then, $c_1 > 0$, $c_2 < 0$, $c_1 + c_2\sigma_1^2 > 0$, $c_1 + c_2\sigma_2^2 < 0$, and

$$S_n = c_1 n + c_2\sum_{k=1}^{n}\frac{(x_{k+1}-x_k)^2}{\eta}.$$

Description of the test: Choose $A_0 > 0$ and $B_0 > 0$. Define $n^* = \min\{n : \ S_n \notin (-B_0, A_0)\}$. We accept $\mathcal{H}_1$ if $S_{n^*} \geq A_0$ and accept $\mathcal{H}_2$ if $S_{n^*} \leq -B_0$.

Proposition 3.2: Under Assumptions (H1) and (H2), there exist a_0 and $\eta_0 > 0$ such that $\min\{A_0, B_0\} \geq a_0|\log\eta|$, for $0 < \eta \leq \eta_0$ imply the error probabilities

$$P(S_{n^*} \leq -B_0|\mathcal{H}_1) \leq \eta \text{ and } P(S_{n^*} \geq A_0|\mathcal{H}_2) \leq \eta.$$

Proof: We only prove the first inequality. The proof for the second one is similar. Assume $\mathcal{H}_1$ holds. Let $\alpha' = (c_1 + c_2\sigma_1^2)/2 > 0$. Then, we have

$$\begin{aligned} P(S_n \leq \alpha' n|\mathcal{H}_1) &= P(c_1 + c_2 Z_n \leq \alpha'|\mathcal{H}_1) \\ &= P(Z_n - \sigma_1^2 \geq \alpha|\mathcal{H}_1) \end{aligned}$$

where

$$\alpha = \frac{(\alpha' - c_1)}{c_2} = -\frac{(c_1 + c_2\sigma_1^2)}{(2c_2)} > 0.$$

Following the proof of Proposition 3.1, we can show there exist k_1 and $\eta_1 > 0$ such that for $0 < \eta < \eta_1$, $n \geq k_1|\log\eta|$

$$P(Z_n - \sigma_1^2 \geq \alpha|\mathcal{H}_1) \leq \eta^3.$$

Take $N_0 = k_1|\log\eta|$. Recall $N_T = T/\eta$. It follows that:

$$\begin{aligned} &P(S_{n^*} \leq -B_0, n^* \geq N_0|\mathcal{H}_1) \\ &\leq \sum_{n=N_0}^{N_T} P(S_n \leq -B_0|\mathcal{H}_1) \\ &\leq \sum_{n=N_0}^{N_T} P(S_n \leq \alpha' n|\mathcal{H}_1) \\ &\leq T\eta^2. \end{aligned}$$

Note that

$$\begin{aligned}&P(S_{n^*} \le -B_0|\mathcal{H}_1)\\ &\quad= P(S_{n^*} \le -B_0, n^* \ge N_0|\mathcal{H}_1)\\ &\quad\quad+ P(S_{n^*} \le -B_0, n^* < N_0|\mathcal{H}_1).\end{aligned}$$

Note also that there exists C_0 such that

$$|S_n| \le C_0 \left(n + 1 + \sup_{k \le N_T} x_k^2 + \sum_{k=1}^{n} w_k^2 \right).$$

Let $\xi_n = B_0/C_0 - n - 1$. Then, for $n < N_0$

$$\begin{aligned}P(|S_n| \ge B_0|\mathcal{H}_1) &\le P\left(\sup_{k \le N_T} x_k^2 \ge \frac{\xi_n}{2} \Big| \mathcal{H}_1 \right)\\ &\quad+ P\left(\sum_{k=1}^{n} w_k^2 \ge \frac{\xi_n}{2} \Big| \mathcal{H}_1 \right).\end{aligned}$$

In view of [14, Lemma 2.1], we can show there exists $\delta > 0$ such that $Ee^{\delta w_k^2} < \infty$ and $Ee^{\delta \sup_{k \le N_T} x_k^2} < \infty$. Therefore, there exist C_1 and C_2 such that

$$P(|S_n| \ge B_0|\mathcal{H}_1) \le C_1 e^{-\delta \xi_n/2} + C_2 e^{-\delta \xi_n/2 + b_0 n}$$

where $b_0 = \log Ee^{\delta w_k^2}$. It follows that there exist $a_0 > 0$, $\eta_1 > 0$ such that for $B_0 \ge a_0|\log \eta|$, $0 < \eta < \eta_0$, and $N < N_0$, we have

$$P(|S_n| \ge B_0|\mathcal{H}_1) \le (C_1 + C_2)\eta^3.$$

Therefore, we have

$$\begin{aligned}&P(S_{n^*} \le -B_0, n^* < N_0|\mathcal{H}_1)\\ &\quad\le \sum_{n=0}^{N_0-1} P(S_n \le -B_0|\mathcal{H}_1)\\ &\quad\le \sum_{n=0}^{N_0-1} P(|S_n| \ge B_0|\mathcal{H}_1)\\ &\quad\le N_0(C_1 + C_2)\eta^3\\ &\quad\le C_3\eta^2\end{aligned}$$

for some C_3. Hence, there exists $\eta_0 > 0$, such that

$$P(S_{n^*} \le -B_0|\mathcal{H}_1) \le \eta$$

for $0 < \eta < \eta_0$ □

Again, this proposition provide a lower bound for A_0 and B_0 to ensure the error probabilities to be smaller than η. More detailed estimates for A_0 and B_0 are used in the numerical section.

IV. Numerical Examples

In this section, we first compare our deep learning-based approach with that of hypothesis test types.

Fixed sample size tests

For the fixed sample tests, we set hypothesis error allowance to be 5%, which leads to the sample length $N^* = \max\{N_1, N_2\}$, where $N_i = 2 \times 1.65^2 \Sigma_0 (1 + 2/\sigma_i^4)$, for $i = 1, 2$, and $\Sigma_0 = ((\sigma_1^2 + \sigma_2^2)/(\sigma_1^2 - \sigma_2^2))^2$. See [14] for details.

TABLE III
Dependence of Accuracy on σ_1 With Fixed $\sigma_2 = 0.5$

σ_1	0.1	0.2	0.3	0.4
N^*	7	12	30	100
Fixed Sample	99.47%	97.95%	97.29%	94.02%
RNN	98.80%	97.37%	96.20%	91.33%

TABLE IV
Dependence of Accuracy on σ_2 With Fixed $\sigma_1 = 0.3$

σ_2	0.4	0.5	0.6	0.7
N^*	80	30	18	14
Fixed Sample	96.09%	97.56%	97.26%	98.04%
RNN	95.43%	96.43%	96.70%	97.47%

TABLE V
Dependence of Accuracy on σ_1 With Fixed $\sigma_2 = 0.5$

σ_1	0.1	0.2	0.3	0.4
(A_0, B_0)	(13.05,0.99)	(8.82,1.86)	(5.92,2.70)	(4.30,3.40)
$\overline{N}$	12	18	31	97
Seq. Test	87.74%	85.84%	81.76%	76.90%
RNN	99.53%	98.63%	97.10%	91.30%

TABLE VI
Dependence of Accuracy on σ_2 With Fixed $\sigma_1 = 0.3$

σ_2	0.4	0.5	0.6	0.7
(A_0, B_0)	(4.61,3.26)	(5.92,2.70)	(7.19,2.55)	(8.30,1.94)
$\overline{N}$	65	30	22	18
Seq. Test	93.60%	96.08%	96.23%	96.79%
RNN	93.00%	96.57%	97.50%	98.03%

Then, we use this sample length for the RNN estimator. We take $b_1(x) = b_2(x) = x$ and $\eta = 0.01$ and vary σ_i. First, we fix $\sigma_2 = 0.5$ and vary σ_1. Then, we fix $\sigma_1 = 0.3$ and vary σ_2. The results are given in Tables III and IV, respectively.

As can be seen from these tables, our RNN estimator performs comparably with the fixed sample size tests.

Sequential tests

Next, for the sequential tests, with hypothesis error 5%, we can calculate the exit interval (A_0, B_0). In particular, A_0 and B_0 are chosen so that the sum of error probabilities $S_{\text{error}} \le 0.05$, where

$$S_{\text{error}} = \frac{1 - e^{-\eta_1 A_0}}{e^{\eta_1 B_0} - e^{-\eta_1 A_0}} + \frac{1 - e^{\eta_2 B_0}}{e^{-\eta_2 A_0} - e^{\eta_2 B_0}}$$

and $\eta_i = (c_1 + c_2\sigma_i^2)/(c_2^2(\sigma_i^4 + 2))$, for $i = 1, 2$; see [14] for details. The average sample lengths $(\overline{N})$ are used for the corresponding RNN sample lengths. Again, we take $b_1(x) = b_2(x) = x$ and $\eta = 0.01$ and vary σ_i. First, we fix $\sigma_2 = 0.5$ and vary σ_1. Then, we fix $\sigma_1 = 0.3$ and vary σ_2. The results are given in Tables V and VI, respectively.

Similarly as in the fixed sample tests, our RNN estimator performs comparably with the sequential tests.

TABLE VII
Mean Reversion Model

η	RNN	FST	N^*	ST	(A_0, B_0)
0.01	90.43%	68.23%	100	69.91%	(2.0,2.0)
0.001	96.80%	77.25%	1000	63.90%	(1.0,1.0)

Mean Reverting Square Root Ornstein-Uhlenbeck Model

We consider the observation data generated with models

$$dX_t = \beta_\theta(\mu_\theta - X_t)dt + \sigma\sqrt{X_t}dV_t,\ \theta \in \{1, 2\}. \tag{44}$$

Its discrete time version is given by

$$x_{k+1} = x_k + \eta\beta_\theta(\mu_\theta - x_k) + \sqrt{\eta}\sigma\sqrt{x_k}w_k.$$

For the numerical simulation, we take $\beta_1 = 1.5, \mu_1 = 0.5$; $\beta_2 = 2.5, \mu_2 = 1.5$; $\sigma = 5.5, T = 1.0, \eta = 0.01, 0.001$. We trained the RNN with 10 000 samples with batch size 1024 for 150 epochs.

In Table VII, FST stands for fixed sample test and ST for sequential test. For the mean reverting square root Ornstein-Uhlenbeck model, considering the nonlinearity of the diffusion term, we employ a grid search scheme to obtain the optimal pair (A_0, B_0) that achieves highest accuracy probability. For the efficiency of the ST scheme, we set $A_0 = B_0$ during grid search and the search range was chosen as $[0.1, 5.0]$. As the choice of N^*, we simply use the entire trajectory of the observations for FST. As can be seen from Table VII, the deep learning-based approach produced the same level performance as in the linear case. On the other hand, note that Assumption (H1) no longer holds. As a result, both the hypothesis tests underperform the deep learning-based approach by a large margin. This example underlines the robustness of our deep NN method.

V. Conclusion

In contrast to the majority of work on system identification for systems given by input and output sequences, this work focused on systems given in continuous time, presented by stochastic differential equations. Our main effort is devoted to computational issues. In particular, we studied two computational methods for system identification, which are the deep learned approach and the hypothesis test methods. We have demonstrated that both of these methods are effective for linear systems. The deep learning-based approach outperforms that of a mean reversion model. This article reports some of our initial findings on system identification. It would be interesting to examine a broader class of linear and nonlinear models. It would be interesting also to extend our results to incorporate models with time dependent unknown parameters or time varying parameters. Such study will be beneficial for numerous systems in a wide range of applications.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, vol. 97, pp. 242–252.

[2] Z. Allen-Zhu, Y. Li, and Z. Song, "On the convergence rate of training recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6676–6688.

[3] G. Battistelli and P. Tesi, "Classification for dynamical systems: Model-based approach and support vector machines," 2018, *arXiv:1803.10552*.

[4] S. Billings, "Identification of nonlinear systems–A survey," *Proc. IEE, Part D*, vol. 127, pp. 272–285, 1980.

[5] P. Billingsley, *Convergence of Probability Measures*. New York, NY, USA: Wiley, 1968.

[6] Huy N. Chau et al., "On fixed gain recursive estimators with discontinuity in the parameters," *ESAIM: Probability Statist.*, vol. 23, pp. 217–244, 2019.

[7] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*. Cambridge, MA, USA: Birkhäuser, 1991.

[8] H. F. Chen and W. Zhao, *Recursive Identification and Parameter Estimation*. Boca Raton, FL, USA: CRC Press, 2014.

[9] A. Daniely, "SGD learns the conjugate Kernel class of the network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 2422–2430.

[10] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. New York, NY, USA: Wiley, 1986.

[11] K. S. Fu, *Syntactic Methods in Pattern Recognition* (Math. Sci. Eng.), vol. 112. New York, NY, USA: Academic, 1974.

[12] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubi, *Bayesian Data Analysi*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2013.

[13] J. Guo and Y. Zhao, "Identification for Wiener-Hammerstein systems under quantized inputs and quantized output observations," *Asian J. Control*, vol. 23, no. 1, pp. 118–127, 2021.

[14] U. G. Haussmann and Q. Zhang, "Discrete time stochastic adaptive control with small observation noise," *Appl. Math. Optim.*, vol. 25, pp. 303–330, 1992.

[15] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York, NY, USA: Wiley, 1996.

[16] Q. He, L. Y. Wang, and G. Yin, *System Identification Using Regular and Quantized Observations: Applications of Large Deviations Principles* (SpringerBriefs in Mathematics). New York, NY, USA: Springer, 2013.

[17] D. Higham, M. Roj, X. Mao, Q. S. Song, and G. Yin, "Mean exit times and the multi-level Monte Carlo method," *SIAM/ASA J. Uncertainty Quantification*, vol. 1, pp. 2–18, 2013.

[18] I.A. Ibragimov and R. Z. Hasminskii, *Statistical Estimation Asymptotic Theory, Translated From the Russian by Samuel Kotz Applications of Mathematics*, vol. 16. New York, NY, USA: Springer-Verlag, 1981.

[19] M. Yu Kaniovskii, "Limit distribution of processes of stochastic approximation type when the regression function has several roots," *Dokl. Akad. Nauk SSSR*, vol. 301, pp. 1308–1309, 1988.

[20] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*. New York, NY, USA: Academic, 1975.

[21] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control* (Classics Appl. Math., 75). Philadelphia, PA, USA: SIAM, 2016.

[22] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, With Applications to Stochastic Systems Theory*. Cambridge, MA, USA: MIT Press, 1984.

[23] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York, NY, USA: Springer, 2003.

[24] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall, 1999.

[25] X. Mao, *Stochastic Differential Equations and Applications*, 2nd ed. Chichester, U.K.: Horwood, 2007.

[26] M. Milanese and G. Belforte, "Estimation theory and uncertainty intervals evaluation in the presence of unknown but bounded errors: Linear families of models and estimators," *IEEE Trans. Autom. Control*, vol. 27, no. 2, pp. 408–414, Apr. 1982.

[27] M. Milanese and A. Vicino, "Information-based complexity and non-parametric worst-case system identification," *J. Complexity*, vol. 9, pp. 427–446, 1993.

[28] B. Mu, E.-W. Bai, W. X. Zheng, and Q. Zhu, "A globally consistent nonlinear least squares estimator for identification of nonlinear rational systems," *Automatica*, vol. 77, pp. 322–335, 2017.

[29] B. Mu, L. Ljung, and T. Chen, "When cannot regularization improve the least squares estimate in the kernel-based regularized system identification," *Automatica*, vol. 160, 2024, Art. no. 111442.

[30] N. Nguyen and G. Yin, "Stochastic approximation with discontinuous dynamics, differential inclusions, and applications," *Ann. Appl. Probability*, vol. 33, pp. 780–823, 2023.

[31] M. Nielsen, Neural Networks and Deep Learning. [Online]. Available: http://neuralnetworksanddeeplearning.com

[32] H. Qian, G. Yin, and Q. Zhang, "Deep filtering with adaptive learning rates," *IEEE Trans. Autom. Control*, Vol. 68, no. 6, pp. 3285–3299, Jun. 2023.

[33] T. Soderstrom and P. Stoica, *System Identification*. Upper Saddle River, NJ, USA: Prentice Hall, 1989.

[34] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer-Verlag, 2000.

[35] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks*, 2nd ed. London, U.K.: Springer, 2003.

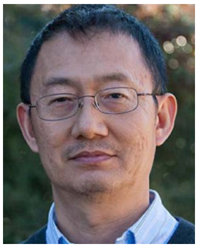
[36] L. Y. Wang, "Persistent identification of time-varying systems," *IEEE Trans. Autom. Control*, vol. 42, no. 1, pp. 66–82, Jan. 1997.

[37] L. Y. Wang, G. Yin, J.-F. Zhang, and Y. L. Zhao, *System Identification With Quantized Observations: Theory and Applications* (Systems & Control: Foundations & Applications). Cambridge, MA, USA: Birkhäuser, 2010.

[38] Y. Wang, Y. Zhao, Ji-Feng Zhang, and J. Guo, "A unified identification algorithm of FIR systems based on binary observations with time-varying thresholds," *Automatica*, vol. 135, 2022, Art. no. 109990.

[39] T. Wigren, "Convergence analysis of recursive identification algorithms based on the nonlinear Wiener model," *IEEE Trans. Autom. Control*, vol. 39, no. 11, pp. 2191–2206, Nov. 1994.

[40] G. Zames, L. Lin, and L. Y. Wang, "Fast identification n-widths and uncertainty principles for LTI and slowly varying systems," *IEEE Trans. Autom. Control*, vol. 39, no. 9, pp. 1827–1838, Sep. 1994.



Qing Zhang (Senior Member, IEEE) received the Ph.D. degree in applied mathematics from Brown University, RI, USA, in 1988.

He has authored or coauthored five monographs on production planning and two-time scale Markovian systems and applications and more than 200 research papers. He has coedited six books. He is currently a Professor of mathematics with the University of Georgia, GA, USA. He specializes in stochastic systems and control, filtering, and applications in finance.

He was Associate Editor for *Automatica*, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and *SIAM Journal on Control and Optimization*. He is currently Corresponding Editor for *SIAM Journal on Control and Optimization*. He also served on a number of international conference organizing committees including Co-Chair of the organizing committee for the SIAM Conference on Control and Applications in 2017.

Xiaohang Ma received the B.S. degree in information and computing science from Huazhong Agricultural University, Wuhan, China, in 2021.

He is currently a Ph.D. candidate with the Department of Mathematics, University of Connecticut, Storrs, CT, USA. His research interests include optimal control, filtering, and optimal stopping for stochastic systems.

George Yin (Life Fellow, IEEE) received the B.S. degree in mathematics from the University of Delaware, Newark, USA, in 1983, and the M.S. degree in electrical engineering and the Ph.D. degree in applied mathematics from Brown University, RI, USA, in 1987.

He joined the Department of Mathematics, Wayne State University, Detroit, MI, USA, in 1987, and became Professor in 1996 and University Distinguished Professor in 2017. He moved to the University of Connecticut, Storrs, CT, USA, in 2020. His research interests include stochastic processes and stochastic systems theory and applications.

Dr. Yin was the Chair of the SIAM Activity Group on Control and Systems Theory, and served on the Board of Directors of the American Automatic Control Council. He was the Editor-in-Chief for *SIAM Journal on Control and Optimization* 2018-2023, Senior Editor for *IEEE Control Systems Letters* 2017–2019, and Associate Editor for *Automatica* 2005–2011, and IEEE TRANSACTIONS ON AUTOMATIC CONTROL 1994–1998. He is a Senior Editor for *Nonlinear Analysis: Hybrid Systems*, and is an Associate Editor for *ESAIM: Control, Optimization and Calculus of Variations*, *Applied Mathematics and Optimization*, and many other journals. He is a Fellow of IFAC and a Fellow of SIAM.