Spectral learning of dynamic systems from nonequilibrium data*

Hao Wu and Frank Noé
Department of Mathematics and Computer Science
Freie Universität Berlin
Arnimallee 6, 14195 Berlin
{hao.wu,frank.noe}@fu-berlin.de

Abstract

Observable operator models (OOMs) and related models are one of the most important and powerful tools for modeling and analyzing stochastic systems. They can exactly describe dynamics of finite-rank systems, and be efficiently learned from data by moment based algorithms. Almost all OOM learning algorithms are developed based on the assumption of equilibrium data which is very difficult to guarantee in real life, especially for complex processes with large time scales. In this paper, we derive a nonequilibrium learning algorithm for OOMs, which dismisses this assumption and can effectively extract the equilibrium dynamics of a system from nonequilibrium observation data. In addition, we propose binless OOMs for the application of nonequilibrium learning to continuous-valued systems. In comparison with the other OOMs with continuous observations, binless OOMs can achieve consistent estimation from nonequilibrium data with only linear computational complexity.

1 Introduction

In the last two decades, a collection of highly related dynamical models including observable operator models (OOMs) [1–3], predictive state representations [4–6] and spectral learning based hidden Markov models [7, 8], have become powerful and increasingly popular tools for analysis of dynamical data. These models are largely similar and can be unified in a general learning framework of multiplicity automata, or equivalently sequential systems [9, 10]. We focus in this paper only on stochastic systems without control inputs. Because all of above mentioned models can be expressed in the form of OOMs for such systems, we will refer to them as OOMs below.

In contrast with the other commonly used models such as Markov models [11], Langevin models [12], traditional hidden Markov models (HMMs) [13] and recurrent neural networks [14], OOMs can exactly characterize the dynamics of a stochastic system without any a priori knowledge except the assumption of finite dynamical rank (i.e., the rank of Hankel matrix) [10], and the parameter estimation can be efficiently performed by the method of moments for discrete-valued systems without solving any intractable inverse or optimization problem.

A major challenge for OOM based dynamical modeling approaches arises from nonequilibrium data. In most literature, the observation data are assumed to be equilibrium so that the expected values of observables associated with OOM learning can be reliably computed by simple averaging. However, the equilibrium assumption can be approximately satisfied only if most of observation data are generated after the system has mixed. In many practical situations, especially where metastable physical or chemical processes are involved, this assumption can be severely violated due to the

*Original title: Learning Observable Operator Models from Nonequilibrium Data

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

limit of experimental technique or computational capacity. A notable example is the distributed computing project Folding@home [15], which explores protein folding processes that occur on the timescales of microseconds to milliseconds based on molecular dynamics simulations on the order of nanoseconds in length. In such a case, it is still unknown how to obtain promising estimates of OOMs from nonequilibrium data consisting of short trajectories. In [16], a hybrid estimation algorithm was proposed to improve OOM learning of large-time-scale processes by using both dynamic and static data, but it still requires assumption of equilibrium data. One solution to reduce the statistical bias caused by nonequilibrium data is to discard the observation data generated before the system reaches steady state, which is a common trick in applied statistics [17]. Obviously, this way suffers from substantial information loss and is infeasible when observation trajectories are shorter than mixing times. Another possible way would be to learn OOMs by likelihood-based estimation instead of moment-based estimation, but there is no effective maximum likelihood or Bayesian estimator of OOMs until now. The maximum pseudo-likelihood estimator of OOMs proposed in [18] demands high computational cost and its consistency is yet unverified.

Another difficulty for OOM based modeling approaches is learning with continuous data, where density estimation problems are involved. The density estimation can be performed by parametric methods such as the fuzzy interpolation [19] and the kernel density estimation [8]. But these methods would reduce the flexibility of OOMs for dynamical modeling because of their limited expressive capacity. Recently, a kernel embedding based OOM learning algorithm was proposed to cope with continuous data [20], which avoids explicit density estimation and learns OOMs in a nonparametric manner. However, the kernel embedding usually yields a very large computational complexity, which greatly limits practical applications of this algorithm to real-world systems.

The purpose of this paper is to address the challenge of nonequilibrium learning of OOMs due to the requirements of analysis of both discrete- and continuous-valued systems. We provide a modified moment-based method for discrete-valued stochastic systems which allows us to consistently estimate the equilibrium dynamics from nonequilibrium data, and then extend this method to OOM learning with continuous observations in a binless manner. In comparison with the existing learning methods for continuous OOMs, the proposed binless method does not rely on any density estimator, and can achieve consistent estimation with linear computational complexity in data size even if the equilibrium assumption of observations does not hold. Moreover, some numerical experiments are provided to demonstrate the capability of the proposed nonequilibrium learning methods.

2 Preliminaries

2.1 Notation

In this paper, we use $\mathbb{P}$ to denote probability distribution for discrete random variables and probability density for continuous random variables. The indicator function of event e is denoted by 1_e and the dirac delta function centered at x is denoted by $\delta_x(\cdot)$. For a given process $\{a_t\}$, we write the subsequence $(a_k, a_{k+1}, \ldots, a_{k'})$ as $a_{k:k'}$, and $\mathbb{E}_\infty[a_t] \triangleq \lim_{t\to\infty} \mathbb{E}[a_t]$ means the expected value of a_t in equilibrium if the limit exists. In addition, the convergence in probability is denoted by $\xrightarrow{p}$.

2.2 Observable operator models

An m-dimensional observable operator model (OOM) with observation space $\mathcal{O}$ can be represented by a tuple $\mathcal{M} = (\boldsymbol{\omega}, \{\boldsymbol{\Xi}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma})$, which consists of an initial state vector $\boldsymbol{\omega} \in \mathbb{R}^{1\times m}$, an evaluation vector $\boldsymbol{\sigma} \in \mathbb{R}^{m\times 1}$ and an observable operator matrix $\boldsymbol{\Xi}(x) \in \mathbb{R}^{m\times m}$ associated to each element $x \in \mathcal{O}$. $\mathcal{M}$ defines a stochastic process $\{x_t\}$ in $\mathcal{O}$ as

$$\mathbb{P}(x_{1:t}|\mathcal{M}) = \boldsymbol{\omega}\boldsymbol{\Xi}(x_{1:t})\boldsymbol{\sigma} \tag{1}$$

with $\boldsymbol{\Xi}(x_{1:t}) \triangleq \boldsymbol{\Xi}(x_1)\ldots\boldsymbol{\Xi}(x_t)$. It is interesting to note that (1) can also be represented in the form of state space models as

$$\begin{aligned} \boldsymbol{\omega}_t &= (\boldsymbol{\omega}_{t-1}\boldsymbol{\Xi}(x_t)\boldsymbol{\sigma})^{-1}\boldsymbol{\omega}_{t-1}\boldsymbol{\Xi}(x_t) \\ \mathbb{P}(x_{t+1}|\boldsymbol{\omega}_t) &= \boldsymbol{\omega}_t\boldsymbol{\Xi}(x_{t+1})\boldsymbol{\sigma} \end{aligned} \tag{2}$$

Here the internal state $\boldsymbol{\omega}_t$ in (2) is a sufficient statistics the process at each time t, which contains all the information needed to predict the future observations and is initialized by $\boldsymbol{\omega}_0 = \boldsymbol{\omega}$. It is clear that two OOMs $\mathcal{M}$ and $\mathcal{M}'$ are equivalent if and only if $\mathbb{P}(x_{1:t}|\mathcal{M}) \equiv \mathbb{P}(x_{1:t}|\mathcal{M}')$.

Algorithm 1 General procedure for OOM learning

INPUT: Observation trajectories generated by a stochastic process $\{x_t\}$ in $\mathcal{O}$

OUTPUT: $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\boldsymbol{\Xi}}(x)\}_{x\in\mathcal{O}}, \hat{\boldsymbol{\sigma}})$

PARAMETER: m: dimension of the OOM. D_1, D_2: numbers of feature functions. L: order of feature functions.

1: Construct feature functions $\boldsymbol{\phi}_1 = (\varphi_{1,1}, \ldots, \varphi_{1,D_1})^\top$ and $\boldsymbol{\phi}_2 = (\varphi_{2,1}, \ldots, \varphi_{2,D_2})^\top$, where each $\varphi_{i,j}$ is a mapping from $\mathcal{O}^L$ to $\mathbb{R}$ and $D_1, D_2 \geq m$.

2: Approximate

$$\bar{\boldsymbol{\phi}}_1 \triangleq \mathbb{E}_\infty\left[\boldsymbol{\phi}_1(x_{t+1:t+L})\right], \quad \bar{\boldsymbol{\phi}}_2 \triangleq \mathbb{E}_\infty\left[\boldsymbol{\phi}_2(x_{t+1:t+L})\right] \tag{6}$$

$$\mathbf{C}_{1,2} \triangleq \mathbb{E}_\infty\left[\boldsymbol{\phi}_1(x_{t-L:t-1})\boldsymbol{\phi}_2(x_{t:t+L-1})^\top\right] \tag{7}$$

$$\mathbf{C}_{1,3}(x) \triangleq \mathbb{E}_\infty\left[1_{x_t=x}\cdot\boldsymbol{\phi}_1(x_{t-L:t-1})\boldsymbol{\phi}_2(x_{t+1:t+L})^\top\right], \quad \forall x \in \mathcal{O} \tag{8}$$

by their empirical means $\hat{\bar{\boldsymbol{\phi}}}_1$, $\hat{\bar{\boldsymbol{\phi}}}_2$, $\hat{\mathbf{C}}_{1,2}$ and $\hat{\mathbf{C}}_{1,3}(x)$ over observation data.

3: Choose matrix $\mathbf{F}_1 \in \mathbb{R}^{D_1\times m}, \mathbf{F}_2 \in \mathbb{R}^{D_2\times m}$ such that $\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,2}\mathbf{F}_2$ is invertible.

4: Compute

$$\hat{\boldsymbol{\sigma}} = \left(\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,2}\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\hat{\bar{\boldsymbol{\phi}}}_1 \tag{9}$$

$$\hat{\boldsymbol{\Xi}}(x) = \left(\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,2}\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,3}(x)\mathbf{F}_2, \quad \forall x \in \mathcal{O} \tag{10}$$

$$\hat{\boldsymbol{\omega}} = \hat{\bar{\boldsymbol{\phi}}}_2^\top\mathbf{F}_2 \tag{11}$$

3 Learning OOMs using moments

3.1 Algorithm

Here and hereafter, we only consider the case that the observation space $\mathcal{O}$ is a finite set. (Learning with continuous observations will be discussed in Section 5.) A large number of largely similar methods have been developed to learn OOMs from discrete data, and the generic learning procedure of these methods is summarized in Algorithm 1 by omitting details of algorithm implementation and parameter choice. For convenience of description and analysis, we specify in this paper the formula for calculating $\hat{\bar{\boldsymbol{\phi}}}_1$, $\hat{\bar{\boldsymbol{\phi}}}_2$, $\hat{\mathbf{C}}_{1,2}$ and $\hat{\mathbf{C}}_{1,3}(x)$ in Line 2 of Algorithm 1 as follows:

$$\hat{\bar{\boldsymbol{\phi}}}_1 \triangleq \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_1(\vec{s}_n^1), \quad \hat{\bar{\boldsymbol{\phi}}}_2 \triangleq \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_1(\vec{s}_n^2) \tag{3}$$

$$\hat{\mathbf{C}}_{1,2} \triangleq \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_1(\vec{s}_n^1)\boldsymbol{\phi}_2(\vec{s}_n^2)^\top \tag{4}$$

$$\hat{\mathbf{C}}_{1,3}(x) \triangleq \frac{1}{N}\sum_{n=1}^{N}1_{s_n^2=x}\boldsymbol{\phi}_1(\vec{s}_n^1)\boldsymbol{\phi}_2(\vec{s}_n^3)^\top, \quad \forall x \in \mathcal{O} \tag{5}$$

Here $\{(\vec{s}_n^1, s_n^2, \vec{s}_n^3)\}_{s=1}^N$ is the collection of all subsequences of length $(2L+1)$ appearing in observation data ($N = T - 2L$ for a single observation trajectory of length T). For instance, if an observation subsequence $x_{t-L:t+L}$ is denoted by $(\vec{s}_n^1, s_n^2, \vec{s}_n^3)$ with some n, then $\vec{s}_n^1 = x_{t-L:t-1}$ and $\vec{s}_n^3 = x_{t+1:t+L}$ represents the prefix and suffix of $x_{t-L:t+L}$ of length L, $s_n^2 = x_t$ is the intermediate observation value, and $\vec{s}_n^2 = x_{t:t+L-1}$ is an "intermediate part" of the subsequence of length L starting from time t (see Fig. 1 for a graphical illustration).

Algorithm 1 is much more efficient than the commonly used likelihood-based learning algorithms and does not suffer from local optima issues. In addition, and more importantly, this algorithm can be shown to be consistent if the observation data are equilibrium so that empirical estimates of $\bar{\boldsymbol{\phi}}_1$, $\bar{\boldsymbol{\phi}}_2$, $\mathbf{C}_{1,2}$ and $\mathbf{C}_{1,3}(x)$ converge to their true values with increasing data size (see, e.g., [3, 8, 10] for

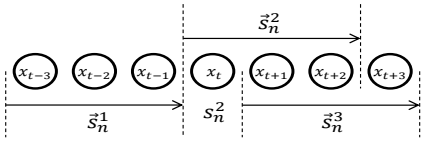

Figure 1: Illustration of variables $\vec{s}_n^1$, s_n^2, $\vec{s}_n^3$ and $\vec{s}_n^2$ used in Eqs. (3)-(5) with $(\vec{s}_n^1, s_n^2, \vec{s}_n^3) = x_{t-L:t+L}$ and $L = 3$.

related works). However, the consistent estimation of OOMs with nonequilibrium data is still an unsolved problem.

3.2 Theoretical analysis

We now analyze statistical properties of the OOM learning algorithm without the assumption of equilibrium observations. Before stating our main result, some assumptions on observation data are listed as follows:

Assumption 1. *The observation data consists of I independent trajectories of length T produced by a stochastic process $\{x_t\}$, and the data size tends to infinity with (i) $I \to \infty$ and $T = T_0$ or (ii) $T \to \infty$ and $I = I_0$.*

Assumption 2. *$\{x_t\}$ is driven by an m-dimensional OOM $\mathcal{M} = (\boldsymbol{\omega}, \{\boldsymbol{\Xi}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma})$, and satisfies*

$$\frac{1}{T'}\sum_{t=1}^{T'} f_t \xrightarrow{p} \mathbb{E}_\infty\left[f_t\right] = \mathbb{E}_\infty\left[f_t | x_{1:k}\right] \tag{12}$$

as $T' \to \infty$ for all k, l, $x_{1:k}$ and $f_t = f\left(x_{t:t+l-1}\right)$.

Assumption 3. *The limit of $\mathbf{F}_1^\top \hat{\mathbf{C}}_{1,2}\mathbf{F}_2 \in \mathbb{R}^{m\times m}$ is invertible.*

Notice that we do not assume stationarity of processes as previously done in the literature, and Assumption 2 only states the asymptotic stationarity of $\{x_t\}$. Therefore, estimates of $\bar{\phi}_1$, $\bar{\phi}_2$, $\mathbf{C}_{1,2}$ and $\mathbf{C}_{1,3}\left(x\right)$ obtained from empirical means may not be consistent if lengths of observation trajectories are kept at finite values (i.e., Case (i) in Assumption 1). Assumption 3 ensures that the limit of $\hat{\mathcal{M}}$ given by Algorithm 1 is well defined.

Based on the above assumptions, we have the following theorem concerning the consistency of the OOM learning algorithm (see Appendix A.1 for proof):

Theorem 1. *Under Assumptions 1-3, the estimated OOM $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\boldsymbol{\Xi}}(x)\}_{x\in\mathcal{O}}, \hat{\boldsymbol{\sigma}})$ given by Algorithm 1 satisfies*

$$\hat{\boldsymbol{\sigma}} \xrightarrow{p} \boldsymbol{\sigma}_{\mathrm{eq}}, \quad \hat{\boldsymbol{\Xi}}\left(x\right) \xrightarrow{p} \boldsymbol{\Xi}_{\mathrm{eq}}(x),\ \forall x \in \mathcal{O} \tag{13}$$

where $\mathcal{M}_{\mathrm{eq}} = (\boldsymbol{\omega}_{\mathrm{eq}}, \{\boldsymbol{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ is an m-dimensional OOM equivalent to $\mathcal{M}$.

This theorem is central in this paper, and implies that the moment based learning algorithm can achieve consistent estimation of all parameters of OOMs except initial state vectors even for nonequilibrium data. ($\hat{\boldsymbol{\omega}} \xrightarrow{p} \boldsymbol{\omega}_{\mathrm{eq}}$ does not hold in most cases except when $\{x_t\}$ is stationary. See Appendix A.1 for details). It can be further generalized according to requirements in more complicated situations where, for example, the data set consists of both several long trajectories and many short trajectories or trajectories are not independent from each other. The following two generalizations are particularly worth mentioning due to their importance for practical applications:

1. The i-th observation trajectories is generated by OOM $\mathcal{M} = (\boldsymbol{\omega}^i, \{\boldsymbol{\Xi}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma})$ for $i = 1, \ldots, I$ (i.e., observations are generated with multiple different initial conditions), and the mean value of $\{\boldsymbol{\omega}^i\}_{i=1}^I$ tends to a constant in probability for $I \to \infty$.
2. Matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ are not constant but given by the singular value decomposition of $\hat{\mathbf{C}}_{1,2}$ as in the spectral learning algorithm [21, 7, 22].

We show in Appendix A that the above two generalizations do not affect the consistency of $\hat{\boldsymbol{\Xi}}\left(x\right)$ and $\hat{\boldsymbol{\sigma}}$. In fact, it can be proved by similar proofs that all theoretical conclusions in this paper hold for the two generalizations.

4 Nonequilibrium learning of OOMs

According to the discussion in the previous section, the only remaining problem for learning OOMs from nonequilibrium data is how to estimate initial state vectors. Considering that the purpose of dynamical modeling is to predict properties of the system in equilibrium in many situations, here we only approximate equilibrium values of internal states of OOMs (see below) rather than actual initial state vectors, because the latter depend on initial conditions of data generation and the former are more physically interesting for analysis of equilibrium dynamics.

Given parameters of $\mathcal{M}_{\mathrm{eq}}$ in Theorem 1, the equilibrium value of the internal state is defined as

$$\bar{\boldsymbol{\omega}}_{\mathrm{eq}} = \lim_{t\to\infty} \boldsymbol{\omega}_{\mathrm{eq}} \boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})^t \tag{14}$$

if the limit exists, where $\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O}) = \sum_{x\in\mathcal{O}} \boldsymbol{\Xi}_{\mathrm{eq}}(x)$. Then the equilibrium dynamics of $\{x_t\}$ can be characterized as

$$\lim_{t\to\infty} \mathbb{P}\left(x_{t+1:t+k} = z_{1:k}\right) = \bar{\boldsymbol{\omega}}_{\mathrm{eq}} \boldsymbol{\Xi}_{\mathrm{eq}}(z_{1:k}) \boldsymbol{\sigma}_{\mathrm{eq}} \tag{15}$$

From (14) and (15), we have

$$\begin{cases} \bar{\boldsymbol{\omega}}_{\mathrm{eq}} \boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O}) = \lim_{t\to\infty} \boldsymbol{\omega}_{\mathrm{eq}} \boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})^{t+1} = \bar{\boldsymbol{\omega}}_{\mathrm{eq}} \\ \bar{\boldsymbol{\omega}}_{\mathrm{eq}} \boldsymbol{\sigma}_{\mathrm{eq}} = \lim_{t\to\infty} \sum_{x\in\mathcal{O}} \mathbb{P}\left(x_{t+1} = x\right) = 1 \end{cases} \tag{16}$$

This motivates the following nonequilibrium learning algorithm for OOMs: *Perform Algorithm 1 to get* $\hat{\boldsymbol{\Xi}}(x)$ *and* $\hat{\boldsymbol{\sigma}}$ *and calculate* $\hat{\boldsymbol{\omega}}$ *by a quandratic programming problem*

$$\hat{\boldsymbol{\omega}} = \arg \min_{\mathbf{w}\in\{\mathbf{w}|\mathbf{w}\hat{\boldsymbol{\sigma}}=1\}} \left\| \mathbf{w}\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{w} \right\|^2 \tag{17}$$

(See Appendix A.4 for a closed-form expression of the solution to (17).)

The existence and uniqueness of $\bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ are shown in Appendix A.4, which yield the following theorem:

Theorem 2. *Under Assumptions 1-3, the estimated OOM* $\hat{\mathcal{M}}$ *provided by the nonequilibrium learning algorithm satisfies*

$$\mathbb{P}\left(x_{1:l} = z_{1:l} | \hat{\mathcal{M}}\right) \xrightarrow{p} \lim_{t\to\infty} \mathbb{P}\left(x_{t+1:t+l} = z_{1:l}\right) \tag{18}$$

for all l *and* $z_{1:l}$.

Remark 1. Some OOM learning algorithms for equilibrium data [23] also calculate $\hat{\boldsymbol{\omega}}$ based on (16), where feature functions $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ and matrices $\mathbf{F}_1, \mathbf{F}_2$ are specifically constructed so that $\hat{\boldsymbol{\omega}}\hat{\boldsymbol{\Xi}}(\mathcal{O}) = \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\omega}}\hat{\boldsymbol{\sigma}} = 1$ can be exactly satisfied even if the statistical noise is considered. In comparison with these algorithms, the nonequilibrium learning algorithm does not require such a restriction, and is shown to be applicable to nonequilibrium data.

5 Binless learning of OOMs

We now consider how to learn OOMs from continuous data. In the case of a real observation space $\mathcal{O} \subset \mathbb{R}^d$, $\mathcal{M}$ defines probability densities of paths of $\{x_t\}$ as in (1), and $\mathbf{C}_{1,3}(x)$ becomes a matrix-valued density function $\mathbf{C}_{1,3}(x) = \frac{1}{\mathrm{d}x}\mathbb{E}_\infty\left[1_{x_t\in\mathrm{d}x} \cdot \boldsymbol{\phi}_1(x_{t-L:t-1})\boldsymbol{\phi}_2(x_{t+1:t+L})^\top\right]$ with general feature functions $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ on $\mathbb{R}^d$, which is difficult to approximate for each $x \in \mathcal{O}$. The existing continuous learning algorithms overcome this problem by using parametric methods [19, 8] or kernel embeddings [20], but none of them can achieve consistent estimation with a low computational complexity like discrete learning algorithms even for equilibrium data.

Here we present a binless strategy to perform dynamical modeling with continuous and nonequilibrium data, which simply views each available observation as a discrete probability atom in the observation space and approximates $\mathbf{C}_{1,3}(x)$ by

$$\hat{\mathbf{C}}_{1,3}(x) = \frac{1}{N}\sum_{n=1}^{N} \delta_{s_n^2}(x)\, \boldsymbol{\phi}_1(\vec{s}_n^1)\boldsymbol{\phi}_2(\vec{s}_n^3)^\top \tag{19}$$

instead of (5). Using this strategy, a binless OOM $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\boldsymbol{\Xi}}(x)\}_{x\in\mathcal{O}}, \hat{\boldsymbol{\sigma}})$ with the observable operator matrix $\hat{\boldsymbol{\Xi}}(x) = \sum_{z\in\mathcal{X}} \hat{\mathbf{W}}_z \delta_z(x)$ supported on $\mathcal{X} = \{s_n^2\}_{n=1}^N$ can be constructed by

Algorithm 2 Nonequilibrium learning procedure of Binless OOMs

INPUT: Observation trajectories generated by a stochastic process $\{x_t\}$ in $\mathcal{O} \subset \mathbb{R}^d$

OUTPUT: Binless OOM $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\boldsymbol{\Xi}}(x)\}_{x\in\mathcal{O}}, \hat{\boldsymbol{\sigma}})$

1: Construct feature functions $\boldsymbol{\phi}_1 : \mathbb{R}^{Ld} \mapsto \mathbb{R}^{D_1}$ and $\boldsymbol{\phi}_2 : \mathbb{R}^{Ld} \mapsto \mathbb{R}^{D_2}$ with $D_1, D_2 \geq m$.

2: Calculate $\bar{\boldsymbol{\phi}}_1, \bar{\boldsymbol{\phi}}_2, \mathbf{C}_{1,2}, \mathbf{C}_{1,3}\,(x)$ by (3), (4) and (19).

3: Choose matrix $\mathbf{F}_1 \in \mathbb{R}^{D_1\times m}, \mathbf{F}_2 \in \mathbb{R}^{D_2\times m}$ such that $\mathbf{F}_1^\top \hat{\mathbf{C}}_{1,2}\mathbf{F}_2$ is invertible.

4: Compute $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\Xi}}(x) = \sum_{z\in\mathcal{X}} \hat{\mathbf{W}}_z \delta_z\,(x)$ by (9), (17) and

$$\hat{\mathbf{W}}_{s_n^2} = \frac{1}{N}\left(\mathbf{F}_1^\top \hat{\mathbf{C}}_{1,2}\mathbf{F}_2\right)^{-1} \mathbf{F}_1^\top \boldsymbol{\phi}_1(\vec{s}_n^1)\boldsymbol{\phi}_2(\vec{s}_n^3)^\top \mathbf{F}_2 \tag{21}$$

where $\hat{\boldsymbol{\Xi}}(\mathcal{O}) = \int \mathrm{d}x\, \hat{\boldsymbol{\Xi}}(x) = \sum_{z\in\mathcal{X}} \hat{\mathbf{W}}_z$.

nonequilibrium learning with computational complexity $O\,(N)$ as in Algorithm 2, where feature functions can be selected as splines, radial basis functions or other commonly used activation functions for single-layer neural networks in practice in order to digest adequate dynamical information from observation data. Note the binless strategy can be applied to more general cases where observations are strings, graphs or other structured variables, and is very similar to that used in Monte Carlo integration or nonparametric maximum likelihood estimation [24]. Although we cannot use the binless OOM to evaluate path probability densities of $\{x_t\}$ as in (18), the equilibrium expectation of any observable $g_t = g\,(x_{t+1:t+r})$ of $\{x_t\}$ can be approximated as

$$\begin{aligned} \mathbb{E}_\infty\,[g_t] &\approx \mathbb{E}\left[g_t|\hat{\mathcal{M}}\right] \\ &= \sum_{x_{1:r}\in\mathcal{X}^r} g\,(x_{1:r})\,\hat{\boldsymbol{\omega}}\hat{\mathbf{W}}_{z_1}\ldots\hat{\mathbf{W}}_{z_r}\hat{\boldsymbol{\sigma}} \end{aligned} \tag{20}$$

By adding a technical assumption, our previous result on consistency of nonequilibrium learning of OOMs can extended to the binless case as follows (see Appendix A.4 for proof):

Assumption 4. *The observation space $\mathcal{O}$ is a closed set in $\mathbb{R}^d$ and feature functions $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ are bounded on $\mathcal{O}^L$.*

Theorem 3. *Under Assumptions 1-4, the binless OOM provided by Algorithm 2 satisfies*

$$\mathbb{E}\left[g\,(x_{1:r})\,|\hat{\mathcal{M}}\right] \xrightarrow{p} \mathbb{E}_\infty\,[g\,(x_{t+1:t+r})] \tag{22}$$

(i) for all continuous functions $g : \mathcal{O}^r \mapsto \mathbb{R}$.

(ii) for all bounded and Borel measurable functions $g : \mathcal{O}^r \mapsto \mathbb{R}$, if there exist constants $\bar{\xi}$ and $\underline{\xi}$ so that $\|\boldsymbol{\Xi}\,(x)\| \leq \bar{\xi}$ and $\lim_{t\to\infty} \mathbb{P}\,(x_{t+1:t+r} = z_{1:r}) \geq \underline{\xi}$ for all $x \in \mathcal{O}$ and $z_{1:r} \in \mathcal{O}^r$.

Note that we do not assume the observed dynamics coincides with a parametric model defined by feature functions in Theorem 3. This theorem shows that binless OOMs allow us to consistently and efficiently extract equilibrium histograms, principle components, time-cross correlations, etc., of a dynamical systems from nonequilibrium data, which is important especially for thermodynamic and kinetic analysis in computational physics and chemistry.

Remark 2. The computational complexity of (20) is $O\,(N^r)$, which is unaffordable for large data sets if $r > 1$. In this paper, we focus on estimation of specific observables in the forms of $\mathbb{E}_\infty\,[a\,(x_t)]$ and $\mathbb{E}_\infty\,[a\,(x_t)\,b\,(x_{t+k})]$ by binless OOMs, which only require $O\,(N)$ time. The efficient estimation of $\mathbb{E}_\infty\,[g\,(x_{t+1:t+r})]$ for general g with is outside the scope of this paper and will be dealt with separately in another paper.

6 Applications

In this section, we evaluate our algorithms on two stochastic systems driven by Brownian dynamics and the molecular dynamics of alanine dipeptide, and compare them to several alternatives. The detailed settings of simulations and algorithms are provided in Appendix B.

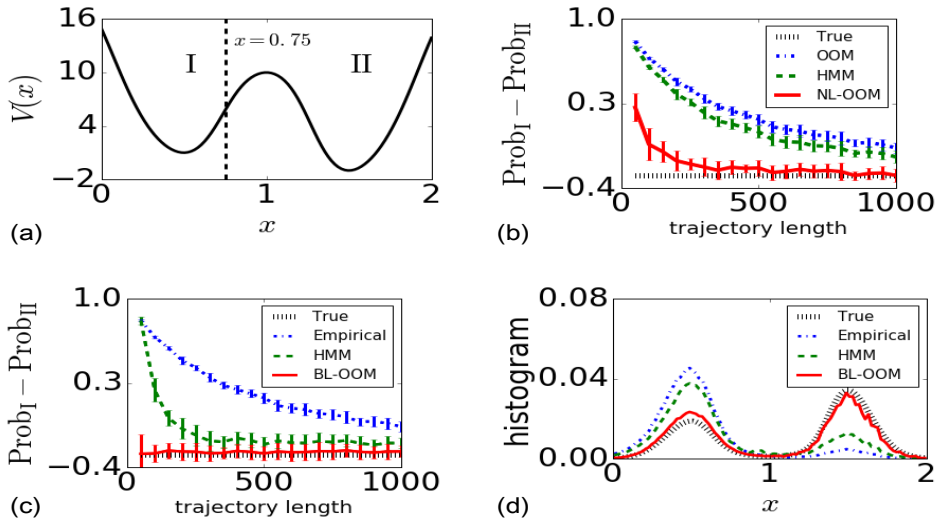

Figure 2: Comparison of modeling methods for a one-dimensional diffusion process. (a) Potential function. (b) Estimates of the difference between equilibrium probabilities of I and II given by the traditional OOM, HMM and OOM using nonequilibrium learning (NL-OOM) with $\mathcal{O} = \{\mathrm{I}, \mathrm{II}\}$. (c) Estimates of the probability difference given by the empirical estimator, HMM and Binless OOM (BL-OOM) using nonequilibrium learning with $\mathcal{O} = [0, 2]$. (d) Equilibrium histograms of $\{x_t\}$ with 100 uniform bins estimated from trajectories with length 50. The initial x_0 are uniformly drawn from $[0, 0.5]$, length of each trajectory is $T = 50 \sim 1000$ and the number of trajectories is $[10^5/T]$. Error bars are standard deviations over 30 independent experiments.

Brownian dynamics Fig. 2(a) shows the potential function of a one-dimensional diffusion process $\{x_t\}$ on $[0, 2]$ driven by Brownian dynamics, where the state space is discretized into two clusters I, II. It is obvious that the equilibrium probability of finding x_t in I is smaller than that of $x_t \in$ II, because the potential well contained in II is deeper than the other one. In this example, all simulations are performed by starting from a uniform distribution on $[0, 0.2]$, which imples that simulations are highly nonequilibrium and it is difficult to accurately estimate the equilibrium probabilities $\mathrm{Prob_I} = \mathbb{E}_\infty [1_{x_t \in \mathrm{I}}]$ and $\mathrm{Prob_{II}} = \mathbb{E}_\infty [1_{x_t \in \mathrm{II}}]$ of I and II from the simulation data. We first utilize the traditional OOM learning, expectation–maximization based HMM learning and the proposed nonequilibrium learning algorithm of OOMs to estimate $\mathrm{Prob_I}$ and $\mathrm{Prob_{II}}$ by assuming that we only know which cluster the x_t is in for each time t, i.e., the observation space $\mathcal{O} = \{\mathrm{I}, \mathrm{II}\}$. Fig. 2(b) summarizes the estimation results with different simulation lengths. It can be seen that estimates given by the traditional OOM and the HMM are far away from true values even for the largest simulation length $T = 1000$. In addition, it is worth pointing out that estimates given by the traditional OOM are very similar to empirical means of $1_{x_t \in \mathrm{I}}$ and $1_{x_t \in \mathrm{II}}$ because the OOM learning algorithm is essentially a moment matching algorithm and the estimated moments cannot be corrected in the traditional learning algorithm. (See Fig. 2(c). Note that the empirical estimates of $\mathrm{Prob_I}$ and $\mathrm{Prob_{II}}$ are the same for discrete and continuous observations.) In contrast to previous methods, the nonequilibrium learning based OOM effectively reduce the statistical bias in the nonequilibrium data, and achieves statistically correct estimation at $T = 300$.

Figs. 2(c) and 2(d) plot estimates of the equilibrium state distribution given by the empirical estimator, HMM and binless OOM using nonequilibrium learning under the condition that the value of x_t is exactly known and $\mathcal{O} = [0, 2]$, where the empirical estimator calculates statistics through averaging over all observations. The observation model of the HMM is constructed based on 100 uniform bins on the state space, where samples within the same bin are assumed to be independent. With such a fine discretization, the performance of the HMM is improved, but estimation errors of the HMM for short trajectory lengths are still large. Here, the proposed binless OOM significantly outperform the other methods, and its estimates are very close to true values even for extremely small short trajectories.

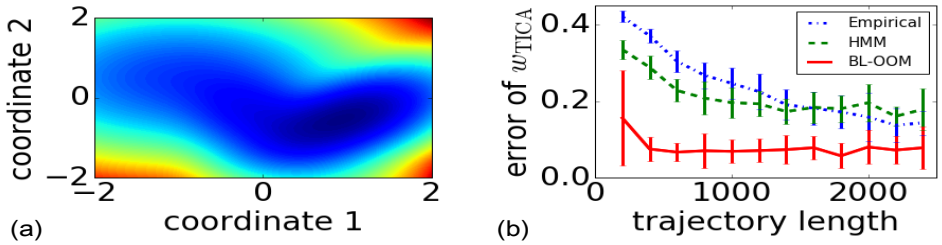

Figure 3: Comparison of modeling methods for a two-dimensional diffusion process. (a) Potential function. (b) Estimates of the coefficient vector $w_{\mathrm{TICA}} \in \mathbb{R}^2$ of the first TIC with lag time 100, which depends on $\mathbb{E}_\infty[x_t]$, $\mathbb{E}_\infty[x_t x_t^\top]$, and $\mathbb{E}_\infty[x_t x_{t+\tau}^\top]$. The initial x_0 are uniformly drawn from $[-2, 0] \times [-2, 0]$, length of each trajectory is $T = 200 \sim 2500$ and the number of trajectories is $[10^5/T]$. Error bars are standard deviations over 30 independent experiments.

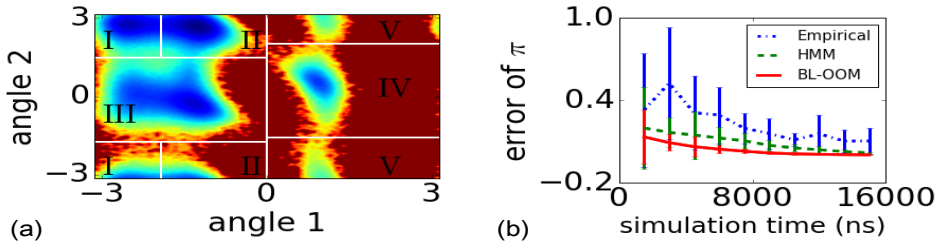

Figure 4: Comparison of modeling methods for molecular dynamics of alanine dipeptide. (a) Reduced free energy. (b) Estimates of $\boldsymbol{\pi}$, the vector of equilibrium probabilities of metastable states I $\sim$ V, where the horizontal axis denotes the total simulation time $T \times I$. Length of each trajectory is $T = 10\mathrm{ns}$ and the number of trajectories is $I = 150 \sim 1500$. Error bars are standard deviations over 30 independent experiments.

Fig. 3 provides an example of applying binless OOMs to kinetic analysis. The goal of this experiment is to perform the time-structure based independent component (TIC) analysis [25] of a two-dimensional Brownian dynamics based on nonequilibrium observation data. Fig. 3(b) displays the estimation errors of the coefficient vector of the first TIC obtained from different learning models, which also demonstrates the superiority of the proposed binless OOM method.

Alanine dipeptide Alanine dipeptide is a small molecule which consists of two alanine amino acid units, and its configuration can be described by two backbone dihedral angles. Fig. 4(a) shows the potential profile of the alanine dipeptide with respect to the two angles, which contains five metastable states. We perform multiple short molecular dynamics simulations starting from the metastable state IV, where each simulation length is $10\mathrm{ns}$, and utilizes different methods to approximate the stationary distribution of the five metastable states. As shown in Fig. 4(b), the proposed binless OOM yields lower estimation error compared to each of the alternatives.

7 Conclusion

In this paper, we investigated the statistical properties of the general OOM learning procedure for nonequilibrium data, and developed a general framework for learning dynamical models from nonequilibrium data. Under this framework, the existing learning approaches of OOMs and the other related models can be conveniently and efficiently applied to nonequilibrium (discrete or continuous) data by using the nonequilibrium learning technique and the binless learning technique. The main ideas of the two techniques are to correct the model parameters by the algebraic constraints under the equilibrium condition and to handle continuous observations in a binless manner. Interesting directions of future research include analysis of approximation error of nonequilibrium learning with finite data size and applications of nonequilibrium learning to controlled systems.

References

[1] H. Jaeger, "Observable operator models for discrete stochastic time series," *Neural Comput.*, vol. 12, no. 6, pp. 1371–1398, 2000.

[2] M.-J. Zhao, H. Jaeger, and M. Thon, "A bound on modeling error in observable operator models and an associated learning algorithm," *Neural Comput.*, vol. 21, no. 9, pp. 2687–2712, 2009.

[3] H. Jaeger, "Discrete-time, discrete-valued observable operator models: a tutorial," tech. rep., International University Bremen, 2012.

[4] M. L. Littman, R. S. Sutton, and S. Singh, "Predictive representations of state," in *Adv. Neural. Inf. Process. Syst. 14 (NIPS 2001)*, pp. 1555–1561, 2001.

[5] S. Singh, M. James, and M. Rudary, "Predictive state representations: A new theory for modeling dynamical systems," in *Proc. 20th Conf. Uncertainty Artif. Intell. (UAI 2004)*, pp. 512–519, 2004.

[6] E. Wiewiora, "Learning predictive representations from a history," in *Proc. 22nd Intl. Conf. on Mach. Learn. (ICML 2005)*, pp. 964–971, 2005.

[7] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," in *Proc. 22nd Conf. Learning Theory (COLT 2009)*, pp. 964–971, 2005.

[8] S. Siddiqi, B. Boots, and G. Gordon, "Reduced-rank hidden Markov models," in *Proc. 13th Intl. Conf. Artif. Intell. Stat. (AISTATS 2010)*, vol. 9, pp. 741–748, 2010.

[9] A. Beimel, F. Bergadano, N. H. Bshouty, E. Kushilevitz, and S. Varricchio, "Learning functions represented as multiplicity automata," *J. ACM*, vol. 47, no. 3, pp. 506–530, 2000.

[10] M. Thon and H. Jaeger, "Links between multiplicity automata, observable operator, models and predictive state representations — a unified learning framework," *J. Mach. Learn. Res.*, vol. 16, pp. 103–147, 2015.

[11] G. R. Bowman, V. S. Pande, and F. Noé, *An introduction to Markov state models and their application to long timescale molecular simulation*. Springer, 2013.

[12] N. Schaudinnus, B. Bastian, R. Hegger, and G. Stock, "Multidimensional langevin modeling of nonoverdamped dynamics," *Phys. Rev. Lett.*, vol. 115, no. 5, p. 050602, 2015.

[13] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comp.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] M. Shirts and V. S. Pande, "Screen savers of the world unite," *Science*, vol. 290, pp. 1903–1904, 2000.

[16] T.-K. Huang and J. Schneider, "Spectral learning of hidden Markov models from dynamic and static data," in *Proc. 30th Intl. Conf. on Mach. Learn. (ICML 2013)*, pp. 630–638, 2013.

[17] M. K. Cowles and B. P. Carlin, "Markov chain monte carlo convergence diagnostics: a comparative review," *J. Am. Stat. Assoc.*, vol. 91, no. 434, pp. 883–904, 1996.

[18] N. Jiang, A. Kulesza, and S. Singh, "Improving predictive state representations via gradient descent," in *Proc. 30th AAAI Conf. Artif. Intell. (AAAI 2016)*, 2016.

[19] H. Jaeger, "Modeling and learning continuous-valued stochastic processes with OOMs," Tech. Rep. GMD-102, German National Research Center for Information Technology (GMD), 2001.

[20] B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola, "Hilbert space embeddings of hidden markov models," in *Proc. 27th Intl. Conf. on Mach. Learn. (ICML 2010)*, 2010.

[21] M. Rosencrantz, G. Gordon, and S. Thrun, "Learning low dimensional predictive representations," in *Proc. 22nd Intl. Conf. on Mach. Learn. (ICML 2004)*, pp. 88–95, ACM, 2004.

[22] B. Boots, *Spectral Approaches to Learning Predictive Representations*. PhD thesis, Carnegie Mellon University, 2012.

[23] M.-J. Zhao, H. Jaeger, and M. Thon, "Making the error-controlling algorithm of observable operator models constructive," *Neural Comput.*, vol. 21, no. 12, pp. 3460–3486, 2009.

[24] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, "A theory of statistical models for monte carlo integration," *J. R. Stat. Soc. B*, vol. 65, no. 3, pp. 585–604, 2003.

[25] G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," *J. Chem. Phys.*, vol. 139, no. 1, p. 015102, 2013.

Supplementary Information

A Proofs

A.1 Proof of Theorem 1

For convenience, here we define

$$\boldsymbol{\omega}'(T) = \frac{1}{T-2L}\boldsymbol{\omega}\sum_{t=1}^{T-2L}\boldsymbol{\Xi}(\mathcal{O})^{t-1}$$

and

$$\mathbf{G}_\sigma = \sum_{z_{1:L}}\boldsymbol{\phi}_2(z_{1:L})\boldsymbol{\sigma}^\top\boldsymbol{\Xi}(z_{1:L})^\top \tag{A.1}$$

Part (1) We first show the theorem in the case of $T = T_0$ and $I \to \infty$.

Let

$$\mathbf{G}_\omega = \sum_{z_{1:L}}\boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}'(T_0)\boldsymbol{\Xi}(z_{1:L}) \tag{A.2}$$

Since $I \to \infty$, we have

$$\begin{aligned}
\hat{\hat{\boldsymbol{\phi}}}_1 &\xrightarrow{p} \mathbb{E}\left[\hat{\hat{\boldsymbol{\phi}}}_1\right] = \mathbf{G}_\omega\boldsymbol{\sigma}\\
\hat{\hat{\boldsymbol{\phi}}}_2^\top &\xrightarrow{p} \mathbb{E}\left[\hat{\hat{\boldsymbol{\phi}}}_2^\top\right] = \boldsymbol{\omega}'(T_0)\mathbf{G}_\sigma^\top\\
\hat{\mathbf{C}}_{1,2} &\xrightarrow{p} \mathbb{E}\left[\hat{\mathbf{C}}_{1,2}\right] = \mathbf{G}_\omega\mathbf{G}_\sigma^\top\\
\hat{\mathbf{C}}_{1,3}(x) &\xrightarrow{p} \mathbb{E}\left[\hat{\mathbf{C}}_{1,2}(x)\right] = \mathbf{G}_\omega\boldsymbol{\Xi}(x)\mathbf{G}_\sigma^\top
\end{aligned}$$

In addition, we can obtain from Assumption 3 that

$$\operatorname{rank}(\mathbf{G}_\omega) = \operatorname{rank}\left(\mathbf{F}_1^\top\mathbf{G}_\omega\right) = \operatorname{rank}(\mathbf{G}_\sigma) = \operatorname{rank}\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right) = m$$

Therefore, $\hat{\mathcal{M}}$ satisfies

$$\begin{aligned}
\hat{\boldsymbol{\omega}} &= \hat{\hat{\boldsymbol{\phi}}}_2^\top\mathbf{F}_2\\
&\xrightarrow{p} \boldsymbol{\omega}'(T_0)\mathbf{G}_\sigma\mathbf{F}_2\\
\hat{\boldsymbol{\Xi}}(x) &= \left(\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,2}\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,3}(x)\mathbf{F}_2\\
&\xrightarrow{p} \left(\mathbf{F}_1^\top\mathbf{G}_\omega\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\mathbf{G}_\omega\boldsymbol{\Xi}(x)\mathbf{G}_\sigma^\top\mathbf{F}_2\\
&= \boldsymbol{\Xi}_{\mathrm{eq}}(x)\\
\hat{\boldsymbol{\sigma}} &= \left(\mathbf{F}_1^\top\hat{\mathbf{C}}_{1,2}\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\hat{\hat{\boldsymbol{\phi}}}_1\\
&\xrightarrow{p} \left(\mathbf{F}_1^\top\mathbf{G}_\omega\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\mathbf{F}_1^\top\mathbf{G}_\omega\boldsymbol{\sigma}\\
&= \boldsymbol{\sigma}_{\mathrm{eq}}
\end{aligned}$$

where $\mathcal{M}_{\mathrm{eq}} = (\boldsymbol{\omega}_{\mathrm{eq}}, \{\boldsymbol{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ is an OOM which equivalent to $\mathcal{M}$ as

$$\begin{aligned}
\boldsymbol{\omega}_{\mathrm{eq}} &= \boldsymbol{\omega}\mathbf{G}_\sigma^\top\mathbf{F}_2\\
\boldsymbol{\Xi}_{\mathrm{eq}}(x) &= \left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\boldsymbol{\Xi}(x)\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right)\\
\boldsymbol{\sigma}_{\mathrm{eq}} &= \left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\boldsymbol{\sigma}
\end{aligned} \tag{A.3}$$

Note $\hat{\boldsymbol{\omega}} \xrightarrow{p} \boldsymbol{\omega}_{\mathrm{eq}}$ does not hold in general cases.

Part (2) We now consider the case of $I = I_0$ and $T \to \infty$.

According to Assumption 2, the limit

$$\begin{aligned}\hat{\mathbf{C}}_{1,2} &\xrightarrow{p} \mathbb{E}_\infty\left[\boldsymbol{\phi}_1(x_{t-L:t-1})\boldsymbol{\phi}_2(x_{t:t+L-1})^\top\right] \\ &= \lim_{k\to\infty}\sum_{z_{1:L}} \boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\boldsymbol{\Xi}(z_{1:L})\mathbf{G}_\sigma^\top\end{aligned}$$

exists. Then

$$\begin{aligned}\hat{\hat{\boldsymbol{\phi}}}_1 &\xrightarrow{p} \mathbb{E}_\infty\left[\boldsymbol{\phi}_1(x_{t-L:t-1})\right] = \mathbf{G}_\omega\boldsymbol{\sigma} \\ \hat{\hat{\boldsymbol{\phi}}}_2^\top &\xrightarrow{p} \mathbb{E}_\infty\left[\boldsymbol{\phi}_2(x_{t:t+L-1})^\top\right] = \lim_{k\to\infty}\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\mathbf{G}_\sigma^\top\mathbf{F}_2 \\ \hat{\mathbf{C}}_{1,2} &\xrightarrow{p} \mathbb{E}_\infty\left[\hat{\mathbf{C}}_{1,2}\right] = \mathbf{G}_\omega\mathbf{G}_\sigma^\top \\ \hat{\mathbf{C}}_{1,3}\left(x\right) &\xrightarrow{p} \mathbb{E}_\infty\left[\hat{\mathbf{C}}_{1,2}\left(x\right)\right] = \mathbf{G}_\omega\boldsymbol{\Xi}(x)\mathbf{G}_\sigma^\top\end{aligned}$$

with

$$\mathbf{G}_\omega = \lim_{k\to\infty}\sum_{z_{1:L}} \boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\boldsymbol{\Xi}(z_{1:L}) \tag{A.4}$$

and we can get

$$\operatorname{rank}\left(\mathbf{G}_\omega\right) = \operatorname{rank}\left(\mathbf{F}_1^\top\mathbf{G}_\omega\right) = \operatorname{rank}\left(\mathbf{G}_\sigma\right) = \operatorname{rank}\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right) = m$$

according to Assumption 3.

Therefore,

$$\begin{aligned}\hat{\boldsymbol{\omega}} &\xrightarrow{p} \lim_{k\to\infty}\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\mathbf{G}_\sigma^\top\mathbf{F}_2 \\ \hat{\boldsymbol{\Xi}}\left(x\right) &\xrightarrow{p} \boldsymbol{\Xi}_{\mathrm{eq}}(x) \\ \hat{\boldsymbol{\sigma}} &\xrightarrow{p} \boldsymbol{\sigma}_{\mathrm{eq}}\end{aligned}$$

where $\mathcal{M}_{\mathrm{eq}} = (\boldsymbol{\omega}_{\mathrm{eq}}, \{\boldsymbol{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ has the same definition as in (A.3). Note $\hat{\boldsymbol{\omega}} \xrightarrow{p} \boldsymbol{\omega}_{\mathrm{eq}}$ does not hold in general cases where $\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right) \neq \boldsymbol{\omega}$.

A.2 Asymptotic correctness of nonequilibrium learning with different initial states

If the i-th observation trajectories is generated by OOM $\mathcal{M} = (\boldsymbol{\omega}^i, \{\boldsymbol{\Xi}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma})$ for $i = 1, \ldots, I$, and

$$\boldsymbol{\omega}'' = \begin{cases} \frac{1}{I}\sum_{i=1}^I \boldsymbol{\omega}^i, & \text{for } T\to\infty \\ \operatorname{plim}_{I\to\infty}\frac{1}{I}\sum_{i=1}^I \boldsymbol{\omega}^i, & \text{for } I\to\infty \end{cases}$$

the asymptotic correctness can also shown as in Appendix A.1 by setting

$$\mathbf{G}_\omega = \sum_{z_{1:L}} \boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}'(T_0)\boldsymbol{\Xi}(z_{1:L})$$

with

$$\boldsymbol{\omega}'(T) = \frac{1}{T-2L}\boldsymbol{\omega}''\sum_{t=1}^{T-2L}\boldsymbol{\Xi}(\mathcal{O})^{t-1}$$

for $I\to\infty$, and

$$\mathbf{G}_\omega = \lim_{k\to\infty}\sum_{z_{1:L}} \boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}''\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\boldsymbol{\Xi}(z_{1:L})$$

for $T\to\infty$.

A.3 Asymptotic correctness of nonequilibrium learning based on the spectral learning algorithm

In the spectral learning algorithm, matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ are given by singular value decomposition (SVD) as

$$\mathbf{F}_1 = \mathbf{U}, \quad \mathbf{F}_2 = \mathbf{V}\mathbf{\Sigma}^{-1} \tag{A.5}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ is a diagonal matrix contains the top m singular values of $\hat{\mathbf{C}}_{1,2}$, and $\mathbf{U}$ and $\mathbf{V}$ consist of the corresponding m left and right singular vectors of $\hat{\mathbf{C}}_{1,2}$. In this appendix, we will prove the following theorem:

Theorem 4. *Under Assumptions 1 and 2, for the estimated OOM $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\mathbf{\Xi}}(x)\}_{x\in\mathcal{O}}, \hat{\boldsymbol{\sigma}})$ given by Algorithm 1 with $\mathbf{F}_1, \mathbf{F}_2$ defined by (A.5), there exists an OOM $\mathcal{M}' = (\boldsymbol{\omega}', \{\mathbf{\Xi}'(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}')$ which is equivalent to $\hat{\mathcal{M}}$ and satisfies*

$$\mathbf{\Xi}'(x) \xrightarrow{p} \mathbf{\Xi}_{\mathrm{eq}}(x), \quad \forall x \in \mathcal{O} \tag{A.6}$$

$$\boldsymbol{\sigma}' \xrightarrow{p} \boldsymbol{\sigma}_{\mathrm{eq}} \tag{A.7}$$

where $\mathcal{M}_{\mathrm{eq}} = (\boldsymbol{\omega}_{\mathrm{eq}}, \{\mathbf{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ is an m-dimensional OOM equivalent to $\mathcal{M}$, if the rank of the limit of $\hat{\mathbf{C}}_{1,2}$ is not less than m.

Proof. According to Appendix A.1, the limit of $\hat{\mathbf{C}}_{1,2}$ can be expressed as

$$\hat{\mathbf{C}}_{1,2} \xrightarrow{p} \mathbf{G}_\omega \mathbf{G}_\sigma^\top$$

where $\mathbf{G}_\omega$ and $\mathbf{G}_\sigma$ have the same definitions as in Appendix A.1. So the limit of $\hat{\mathbf{C}}_{1,2}$ has rank m. By the Eckart-Young-Mirsky Theorem, $\hat{\mathbf{C}}_{1,2}^{\mathrm{trun}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the best rank m approximation to $\hat{\mathbf{C}}_{1,2}$ and therefore

$$\hat{\mathbf{C}}_{1,2}^{\mathrm{trun}} \xrightarrow{p} \mathbf{G}_\omega \mathbf{G}_\sigma^\top$$

Let

$$\mathbf{G}_\omega \mathbf{G}_\sigma^\top = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$$

be the SVD of $\mathbf{G}_\omega \mathbf{G}_\sigma^\top$,

$$\tilde{\mathbf{F}}_1 = \tilde{\mathbf{U}}, \quad \tilde{\mathbf{F}}_2 = \tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-1}$$

and $\mathcal{M}_{\mathrm{eq}} = (\boldsymbol{\omega}_{\mathrm{eq}}, \{\mathbf{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ be an OOM which is equivalent to $\mathcal{M}$ with

$$\begin{aligned}
\boldsymbol{\omega}_{\mathrm{eq}} &= \boldsymbol{\omega}\mathbf{G}_\sigma^\top \tilde{\mathbf{F}}_2 \\
\mathbf{\Xi}_{\mathrm{eq}}(x) &= \left(\mathbf{G}_\sigma^\top \tilde{\mathbf{F}}_2\right)^{-1} \mathbf{\Xi}(x) \left(\mathbf{G}_\sigma^\top \tilde{\mathbf{F}}_2\right) \\
\boldsymbol{\sigma}_{\mathrm{eq}} &= \left(\mathbf{G}_\sigma^\top \tilde{\mathbf{F}}_2\right)^{-1} \boldsymbol{\sigma}
\end{aligned}$$

We can obtain from the Wedin Theorem and the continuity of singular values of matrice that

$$\begin{aligned}
\min_{\mathbf{R}} \left\| \mathbf{U}\mathbf{R} - \tilde{\mathbf{U}} \right\| = \left\| \mathbf{U}\mathbf{U}^\top \tilde{\mathbf{U}} - \tilde{\mathbf{U}} \right\| &\xrightarrow{p} 0 \\
\min_{\mathbf{R}} \left\| \mathbf{V}\mathbf{R} - \tilde{\mathbf{V}} \right\| = \left\| \mathbf{V}\mathbf{V}^\top \tilde{\mathbf{V}} - \tilde{\mathbf{V}} \right\| &\xrightarrow{p} 0 \\
\mathbf{\Sigma} &\xrightarrow{p} \tilde{\mathbf{\Sigma}}
\end{aligned}$$

Therefore, we can construct an OOM $\mathcal{M}' = (\boldsymbol{\omega}', \{\boldsymbol{\Xi}'(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}')$ with

$$\begin{aligned}
\boldsymbol{\omega}' &= \hat{\boldsymbol{\omega}}\left(\boldsymbol{\Sigma}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right) \\
&= \hat{\tilde{\boldsymbol{\phi}}}_2^\top\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1} \\
\boldsymbol{\Xi}'(x) &= \left(\boldsymbol{\Sigma}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\hat{\boldsymbol{\Xi}}(x)\left(\boldsymbol{\Sigma}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right) \\
&= \left(\tilde{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\hat{\mathbf{C}}_{1,2}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\tilde{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\hat{\mathbf{C}}_{1,3}(x)\,\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1} \\
\boldsymbol{\sigma}' &= \left(\boldsymbol{\Sigma}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\hat{\boldsymbol{\sigma}} \\
&= \left(\tilde{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\hat{\mathbf{C}}_{1,2}\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\tilde{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top\hat{\tilde{\boldsymbol{\phi}}}_1
\end{aligned}$$

which is equivalent to $\hat{\mathcal{M}}$ and satisfies

$$\begin{aligned}
\boldsymbol{\Xi}'(x) &\xrightarrow{p} \left(\tilde{\mathbf{U}}^\top\mathbf{G}_\omega\mathbf{G}_\sigma^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\tilde{\mathbf{U}}^\top\mathbf{G}_\omega\boldsymbol{\Xi}(x)\mathbf{G}_\sigma^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1} \\
&= \boldsymbol{\Xi}_{\mathrm{eq}}(x) \\
\boldsymbol{\sigma}' &\xrightarrow{p} \left(\tilde{\mathbf{U}}^\top\mathbf{G}_\omega\mathbf{G}_\sigma^\top\tilde{\mathbf{V}}\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\tilde{\mathbf{U}}^\top\mathbf{G}_\omega\boldsymbol{\sigma} \\
&= \boldsymbol{\sigma}_{\mathrm{eq}}
\end{aligned}$$

□

It is worth pointing out that we can also show conclusions of Theorems 2 and 3 with (A.5) by using similar proofs. The details proofs are omitted because they are trivial.

A.4 Proof of Theorem 2

Part (1) We first show that there is an OOM $\bar{\mathcal{M}}_{\mathrm{eq}} = (\bar{\boldsymbol{\omega}}_{\mathrm{eq}}, \{\boldsymbol{\Xi}_{\mathrm{eq}}(x)\}_{x\in\mathcal{O}}, \boldsymbol{\sigma}_{\mathrm{eq}})$ which can describe the equilibrium dynamics of $\{x_t\}$, where $\boldsymbol{\Xi}_{\mathrm{eq}}(x)$ and $\boldsymbol{\sigma}_{\mathrm{eq}}$ are defined in (A.3).

In the case of $T = T_0$ and $I \to \infty$, we can obtain from Assumptions 2 and 3 that

$$\begin{aligned}
\lim_{k\to\infty}\mathbf{G}_\omega\boldsymbol{\Xi}(\mathcal{O})^k\mathbf{G}_\sigma^\top &= \lim_{k\to\infty}\frac{1}{T_0-2L}\sum_{t=0}^{T_0-2L-1}\mathbb{E}\left[\boldsymbol{\phi}_1\left(x_{t+1:t+L}\right)\boldsymbol{\phi}_2\left(x_{t+L+k+1:t+2L+k}\right)^\top\right] \\
&= \left(\frac{1}{T_0-2L}\sum_{t=0}^{T_0-2L-1}\mathbb{E}\left[\boldsymbol{\phi}_1\left(x_{t+1:t+L}\right)\right]\right)\left(\mathbb{E}_\infty\left[\boldsymbol{\phi}_2\left(x_{t+1:t+L}\right)^\top\right]\right) \\
&= \mathbf{G}_\omega\boldsymbol{\sigma}\left(\mathbb{E}_\infty\left[\boldsymbol{\phi}_2\left(x_{t+1:t+L}\right)^\top\right]\right) \\
\Rightarrow \lim_{k\to\infty}\boldsymbol{\Xi}(\mathcal{O})^k &= \boldsymbol{\sigma}\bar{\boldsymbol{\omega}} \qquad (A.8)
\end{aligned}$$

with

$$\bar{\boldsymbol{\omega}} = \left(\mathbb{E}_\infty\left[\boldsymbol{\phi}_2\left(x_{t+1:t+L}\right)^\top\right]\right)\mathbf{G}_\sigma^{+\top} \qquad (A.9)$$

where $\mathbf{G}_\omega$ and $\mathbf{G}_\sigma$ are defined by (A.2) and (A.1), and $\mathbf{G}_\sigma^+$ denotes the Moore-Penrose pseudoinverse of $\mathbf{G}_\sigma$. Then

$$\begin{aligned}
\lim_{t\to\infty}\mathbb{P}\left(x_{t+1:t+l} = z_{1:l}\right) &= \lim_{t\to\infty}\boldsymbol{\omega}\boldsymbol{\Xi}(\mathcal{O})^t\boldsymbol{\Xi}(z_{1:l})\boldsymbol{\sigma} \\
&= \boldsymbol{\omega}\boldsymbol{\Xi}(\mathcal{O})\boldsymbol{\sigma}\bar{\boldsymbol{\omega}}\boldsymbol{\Xi}(z_{1:l})\boldsymbol{\sigma} \\
&= \bar{\boldsymbol{\omega}}\boldsymbol{\Xi}(z_{1:l})\boldsymbol{\sigma} \\
&= \bar{\boldsymbol{\omega}}_{\mathrm{eq}}\boldsymbol{\Xi}_{\mathrm{eq}}(z_{1:l})\boldsymbol{\sigma}_{\mathrm{eq}}
\end{aligned}$$

with

$$\bar{\boldsymbol{\omega}}_{\mathrm{eq}} = \bar{\boldsymbol{\omega}}\mathbf{G}_\sigma^\top\mathbf{F}_2 \qquad (A.10)$$

In the case of $T = T_0$ and $I \to \infty$, because $\mathrm{rank}\,(\mathbf{G}_\omega) = m$ for $\mathbf{G}_\omega$ defined by (A.4), there is a sufficiently large but finite T' so that $\mathrm{rank}\,(\mathbf{G}'_\omega) = m$ with

$$\mathbf{G}'_\omega = \sum_{z_{1:L}} \boldsymbol{\phi}_1(z_{1:L})\boldsymbol{\omega}\boldsymbol{\Xi}\left(\mathcal{O}\right)^{T'}\boldsymbol{\Xi}(z_{1:L})$$

Considering

$$\begin{aligned}
\lim_{k\to\infty}\mathbf{G}'_\omega\boldsymbol{\Xi}(\mathcal{O})^k\mathbf{G}_\sigma^\top &= \lim_{k\to\infty}\mathbb{E}\left[\boldsymbol{\phi}_1\left(x_{T'+1:T'+L}\right)\boldsymbol{\phi}_2\left(x_{T'+L+k+1:T'+2L+k}\right)^\top\right] \\
&= \mathbf{G}'_\omega\boldsymbol{\sigma}\left(\mathbb{E}_\infty\left[\boldsymbol{\phi}_2\left(x_{t+1:t+L}\right)^\top\right]\right) \\
\Rightarrow \lim_{k\to\infty}\boldsymbol{\Xi}(\mathcal{O})^k &= \boldsymbol{\sigma}\bar{\boldsymbol{\omega}} \qquad\qquad (A.11)
\end{aligned}$$

with $\bar{\boldsymbol{\omega}}$ defined by (A.9), we can also conclude that

$$\lim_{t\to\infty}\mathbb{P}\left(x_{t+1:t+l} = z_{1:l}\right) = \bar{\boldsymbol{\omega}}_{\mathrm{eq}}\boldsymbol{\Xi}_{\mathrm{eq}}(z_{1:l})\boldsymbol{\sigma}_{\mathrm{eq}}$$

with $\bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ defined by (A.10).

Note in both cases, $\bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ satisfies $\boldsymbol{\omega}_{\mathrm{eq}}\lim_{k\to\infty}\boldsymbol{\Xi}(\mathcal{O})^k = \bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ and

$$\begin{aligned}
\bar{\boldsymbol{\omega}}_{\mathrm{eq}}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O}) &= \lim_{t\to\infty}\boldsymbol{\omega}_{\mathrm{eq}}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})^{t+1} \\
&= \bar{\boldsymbol{\omega}}_{\mathrm{eq}} \\
\bar{\boldsymbol{\omega}}_{\mathrm{eq}}\boldsymbol{\sigma}_{\mathrm{eq}} &= \bar{\boldsymbol{\omega}}_{\mathrm{eq}}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})\boldsymbol{\sigma}_{\mathrm{eq}} \\
&= \lim_{t\to\infty}\sum_{x\in\mathcal{O}}\mathbb{P}\left(x_t = x\right) = 1
\end{aligned}$$

Part (2) In this part, we show that

$$\mathbf{w}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O}) = \mathbf{w}, \quad \mathbf{w}\boldsymbol{\sigma}_{\mathrm{eq}} = 1$$

has a unique solution $\mathbf{w} = \bar{\boldsymbol{\omega}}_{\mathrm{eq}}$.

According to Appendix A.1 and (A.8), (A.11), if $\mathbf{w}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O}) = \mathbf{w}$ and $\mathbf{w}\boldsymbol{\sigma}_{\mathrm{eq}} = 1$, we have

$$\begin{aligned}
\mathbf{w} &= \lim_{k\to\infty}\mathbf{w}\boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})^k \\
&= \lim_{k\to\infty}\mathbf{w}\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\boldsymbol{\Xi}(\mathcal{O})^k\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right) \\
&= \mathbf{w}\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right)^{-1}\boldsymbol{\sigma}\bar{\boldsymbol{\omega}}\left(\mathbf{G}_\sigma^\top\mathbf{F}_2\right) \\
&= \mathbf{w}\boldsymbol{\sigma}_{\mathrm{eq}}\bar{\boldsymbol{\omega}}_{\mathrm{eq}} \\
&= \bar{\boldsymbol{\omega}}_{\mathrm{eq}}
\end{aligned}$$

Part (3) We now show Theorem 2.

The optimization problem (17) can be equivalently transformed into an unconstrained problem

$$\hat{\boldsymbol{\omega}} = \min_{\mathbf{w}}\left\|\mathbf{w}^{\mathrm{proj}}\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{w}^{\mathrm{proj}}\right\|^2 + \left\|\mathbf{w}^{\mathrm{proj}} - \mathbf{w}\right\|^2$$

where

$$\mathbf{w}^{\mathrm{proj}} = \mathbf{w}\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^+\right) + \hat{\boldsymbol{\sigma}}^+ \qquad (A.12)$$

denotes the projection of $\mathbf{w}$ on the space $\{\mathbf{w}|\mathbf{w}\hat{\boldsymbol{\sigma}} = 1\}$, $\mathbf{I}$ denotes the identity matrix of appropriate dimension, and $\bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ is the unique solution if $\hat{\boldsymbol{\Xi}}(\mathcal{O}) = \boldsymbol{\Xi}_{\mathrm{eq}}(\mathcal{O})$ and $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}_{\mathrm{eq}}$ according to the conclusion in Part (2). Then we can obtain that $\hat{\boldsymbol{\omega}} \xrightarrow{p} \bar{\boldsymbol{\omega}}_{\mathrm{eq}}$ according to Theorem 2.7 in [1], which yields the conclusion of Theorem 2.

Part (4) We derive in this part the closed-form solution to (17).

Since the projection of $\mathbf{w}$ on the space $\{\mathbf{w}|\mathbf{w}\hat{\boldsymbol{\sigma}} = 1\}$ is $\mathbf{w}^{\text{proj}}$ defined by(A.12), (17) can be equivalent transformed into

$$\min_{\mathbf{w}} \left\| \mathbf{w}\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^{+}\right)\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right) + \hat{\boldsymbol{\sigma}}^{+}\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right)\right\|^2$$

The solution to this problem is

$$\mathbf{w}^{*} = -\hat{\boldsymbol{\sigma}}^{+}\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right)\left(\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^{+}\right)\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right)\right)^{+}$$

which provides the optimal value of $\hat{\boldsymbol{\omega}}$ as

$$\begin{aligned}\hat{\boldsymbol{\omega}} &= \mathbf{w}^{*}\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^{+}\right) + \hat{\boldsymbol{\sigma}}^{+} \\ &= \hat{\boldsymbol{\sigma}}^{+} - \hat{\boldsymbol{\sigma}}^{+}\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right)\left(\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^{+}\right)\left(\hat{\boldsymbol{\Xi}}(\mathcal{O}) - \mathbf{I}\right)\right)^{+}\left(\mathbf{I} - \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^{+}\right) \qquad \text{(A.13)}\end{aligned}$$

A.5 Proof of Theorem 3

Here we only consider the consistency of the binless OOM as $I \to \infty$. The proof can be easily to extended to the case of $T \to \infty$. In addition, we denote $\mathbb{E}_{\infty}[g(x_{t+1:t+r})]$ and $\mathbb{E}[g(x_{1:r})|\hat{\mathcal{M}}]$ by $\mathbb{E}_{\infty}[g]$ and $\mathbb{E}_{\hat{\mathcal{M}}}[g]$ for convenience of notation.

Part (1) We first show that Theorem 3 holds for $g\left(x_{t+1:t+r}\right) = 1_{x_{t+1:t+r}\in\mathcal{B}_{i_1}\times\mathcal{B}_{i_2}\times\ldots\times\mathcal{B}_{i_r}}$, where $\mathcal{B}_1, \ldots, \mathcal{B}_K$ is a partition of $\mathcal{O}$, $i_{1:r} \in \{1, \ldots, K\}^r$, and 1_e denotes the indicator function of event e. In this case, we can construct a discrete OOM with observation space $\{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$ by the nonequilibrium learning algorithm, which can provide the same estimate of $\mathbb{E}_{\infty}\left[g\left(x_{t+1:t+r}\right)\right]$ as $\hat{\mathcal{M}}$. Therefore, we can show $\mathbb{E}_{\hat{\mathcal{M}}}[g] \xrightarrow{p} \mathbb{E}_{\infty}[g]$ by using the similar proof of Theorem 2.

Part (2) We now consider the case that g is a continuous function. According to the Heine-Cantor theorem, g is also uniformly continuous. Then, for an abitrary $\epsilon > 0$, we can construct a simple function

$$\hat{g}(x_{t+1:t+r}) = \sum_{i_1,\ldots,i_r} c_{i_1 i_2 \ldots i_r} 1_{x_{t+1:t+r}\in\mathcal{B}_{i_1}\times\ldots\times\mathcal{B}_{i_r}}$$

so that

$$|g(z_{1:r}) - \hat{g}(z_{1:r})| \leq \epsilon, \quad \forall z_{1:r} \in \mathcal{O}^r$$

where $\{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$ is a partition of $\mathcal{O}$. Then, we have

$$|\mathbb{E}_{\infty}[g] - \mathbb{E}_{\infty}[\hat{g}]| \leq \mathbb{E}_{\infty}[|g - \hat{g}|] \leq \epsilon$$

and

$$\left|\mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right| \xrightarrow{p} 0$$

as $I \to \infty$ according to the conclusion of Part (1), where $\mathbb{E}_{\infty}[g] = \mathbb{E}_{\infty}[g(x_{t+1:t+r})]$ and $\mathbb{E}_{\hat{\mathcal{M}}}[g] = \mathbb{E}[g(x_{1:r})|\hat{\mathcal{M}}]$.

It can be known from Assumption 4, there exists a constant ξ such that

$$1_{\max_{x\in\mathcal{X}}\left\|\hat{\boldsymbol{\Xi}}(x)\right\| < \xi/|\mathcal{X}|} \xrightarrow{p} 1 \qquad \text{(A.14)}$$

Under the condition that $\max_{x\in\mathcal{X}}\left\|\hat{\boldsymbol{\Xi}}(x)\right\| < \xi/|\mathcal{X}|$, we have

$$\begin{aligned}\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| &= \hat{\boldsymbol{\omega}}_0\left(\sum_{z_{1:r}\in\mathcal{X}^r}\left(\hat{g}(z_{1:r}) - g(z_{1:r})\right)\hat{\boldsymbol{\Xi}}(z_{1:r})\right)\hat{\boldsymbol{\sigma}} \\ &\leq \|\hat{\boldsymbol{\omega}}_0\|\,\|\hat{\boldsymbol{\sigma}}\|\left(\sum_{z_{1:r}\in\mathcal{X}^r}\frac{\xi^r\epsilon}{|\mathcal{X}|^r}\right) \\ &= \|\hat{\boldsymbol{\omega}}_0\|\,\|\hat{\boldsymbol{\sigma}}\|\,\xi^r\epsilon\end{aligned}$$

In addition, considering that we can show as in Appendix A.1 that $\hat{\boldsymbol{\omega}} \xrightarrow{p} \bar{\boldsymbol{\omega}}_{\text{eq}}$ and $\hat{\boldsymbol{\sigma}} \xrightarrow{p} \boldsymbol{\sigma}_{\text{eq}}$, we can obtain

$$1_{\|\hat{\boldsymbol{\omega}}\|\|\hat{\boldsymbol{\sigma}}\| \leq \xi_0} \xrightarrow{p} 1 \tag{A.15}$$

and

$$1_{\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \leq \xi_0 \xi^r \epsilon} \xrightarrow{p} 1$$

where ξ_0 is a constant larger than $\|\bar{\boldsymbol{\omega}}_{\text{eq}}\| \cdot \|\boldsymbol{\sigma}_{\text{eq}}\|$.

Based on the above analysis and the fact that

$$\begin{aligned} \left|\mathbb{E}_{\infty}[g] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| &= \left|\mathbb{E}_{\infty}[g] - \mathbb{E}_{\infty}[\hat{g}] + \mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] + \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \\ &\leq |\mathbb{E}_{\infty}[g] - \mathbb{E}_{\infty}[\hat{g}]| + \left|\mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right| + \left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \end{aligned}$$

we can get

$$\begin{aligned} \Pr\left(\left|\mathbb{E}_{\infty}[g] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \leq (\xi_0 \xi^r + 2)\,\epsilon\right) &\geq \Pr\left(|\mathbb{E}_{\infty}[g] - \mathbb{E}_{\infty}[\hat{g}]| \leq \epsilon, \left|\mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right| \leq \epsilon,\right. \\ &\qquad \left.\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \leq \xi_0 \xi^r \epsilon\right) \\ &\to 1 \end{aligned}$$

Because this equation holds for all $\epsilon > 0$, we can conclude that $\mathbb{E}_{\hat{\mathcal{M}}}[g] \xrightarrow{p} \mathbb{E}_{\infty}[g]$.

Part (3) In this part, we prove the conclusion of the theorem in the case where g is a Borel measurable function and bounded with $|g(z_{1:r})| < \xi_g$ for all $z_{1:r} \in \mathcal{O}^r$, and there exist constants $\bar{\xi}$ and $\underline{\xi}$ so that $\|\boldsymbol{\Xi}(x)\| \leq \bar{\xi}$ and $\lim_{t\to\infty} \mathbb{P}(x_{t+1:t+r} = z_{1:r}) \geq \underline{\xi}$ for all $x \in \mathcal{O}$ and $z_{1:r} \in \mathcal{O}^r$.

According to Theorem 2.2 in [2], for an arbitrary $\epsilon > 0$, there is a continuous function $\hat{g}'$ satisfies $\mathbb{E}_{\infty}[1_{x_{t+1:t+r} \in \mathcal{K}_\epsilon(\hat{g}')}] < \epsilon$, where $\mathcal{K}_\epsilon(\hat{g}') = \{z_{1:r} | z_{1:r} \in \mathcal{O}^r, |\hat{g}'(z_{1:r}) - g(z_{1:r})| > \epsilon\}$. Define

$$\hat{g}(z_{1:r}) = \begin{cases} \hat{g}'(z_{1:r}), & |\hat{g}'(z_{1:r})| \leq \xi_g \\ -\xi_g, & \hat{g}'(z_{1:r}) < -\xi_g \\ \xi_g, & \hat{g}'(z_{1:r}) > \xi_g \end{cases}$$

It can be seen that $\hat{g}$ is a continuous function which is also satisfies $\mathbb{E}_{\infty}[1_{x_{t+1:t+r} \in \mathcal{K}_\epsilon(\hat{g})}] < \epsilon$ and bounded with $|\hat{g}(z_{1:r})| < \xi_g$. So the difference between $\mathbb{E}_{\infty}[g]$ and $\mathbb{E}_{\infty}[\hat{g}]$ satisfies

$$\begin{aligned} |\mathbb{E}_{\infty}[g] - \mathbb{E}_{\infty}[\hat{g}]| &\leq \mathbb{E}_{\infty}\left[|g(x_{t+1:t+r}) - \hat{g}(x_{t+1:t+r})|\right] \\ &= \mathbb{E}_{\infty}[1_{x_{t+1:t+r} \in \mathcal{K}_\epsilon(\hat{g})}] \mathbb{E}_{\infty}\left[|g(x_{t+1:t+r}) - \hat{g}(x_{t+1:t+r})|\, | x_{t+1:t+r} \in \mathcal{K}_\epsilon(\hat{g})\right] \\ &\quad + \mathbb{E}_{\infty}[1_{x_{t+1:t+r} \notin \mathcal{K}_\epsilon(\hat{g})}] \mathbb{E}_{\infty}\left[|g(x_{t+1:t+r}) - \hat{g}(x_{t+1:t+r})|\, | x_{t+1:t+r} \notin \mathcal{K}_\epsilon(\hat{g})\right] \\ &\leq \epsilon \cdot 2\xi_g + \epsilon = (2\xi_g + 1)\,\epsilon \end{aligned}$$

For the difference between $\mathbb{E}_{\infty}[\hat{g}]$ and $\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]$, we can obtain from the above that $\left|\mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right| \xrightarrow{p} 0$ as $I \to \infty$ by considering that $\hat{g}$ is continuous, which implies that there is an I_0 such that

$$\Pr\left(\left|\mathbb{E}_{\infty}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right| > \epsilon\right) < \epsilon, \quad \forall I > I_0$$

Next, let us consider the value of $\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right|$. Note that

$$\begin{aligned} \left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| &\leq \|\hat{\boldsymbol{\omega}}_0\|\,\|\hat{\boldsymbol{\sigma}}\| \left\| \sum_{z_{1:n} \in \mathcal{X}^r} (\hat{g}(z_{1:r}) - g(z_{1:r}))\, \hat{\boldsymbol{\Xi}}(z_{1:r}) \right\| \\ &< \frac{\xi_0 \xi^r}{|\mathcal{X}|^r} \left| \sum_{z_{1:r} \in \mathcal{X}^r} (\hat{g}(z_{1:r}) - g(z_{1:r})) \right| \end{aligned}$$

under the condition that $\left\|\hat{\boldsymbol{\Xi}}(x)\right\| < \xi / |\mathcal{X}|$ and $\|\hat{\boldsymbol{\omega}}\|\,\|\hat{\boldsymbol{\sigma}}\| \leq \xi_0$. Therefore, there exists an I_1 such that

$$\Pr\left(\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}] - \mathbb{E}_{\hat{\mathcal{M}}}[g]\right| \geq \frac{\xi_0 \xi^r}{|\mathcal{X}|^r} \left| \sum_{z_{1:r} \in \mathcal{X}^r} (\hat{g}(z_{1:r}) - g(z_{1:r})) \right|\right) < \epsilon, \quad \forall I > I_1 \tag{A.16}$$

due to (A.14) and (A.15). Let $x'_{1:r}$ denotes a random sample taken uniformly from $\mathcal{X}^r$. We can obtain that

$$\begin{aligned}\mathbb{P}\left(x'_{1:r}\right) &= \mathbb{P}\left(x'_1\right)\ldots\mathbb{P}\left(x'_r\right)\\ &\leq \left(\|\boldsymbol{\omega}\|\,\|\boldsymbol{\sigma}\|\,\xi_O\bar{\xi}\right)^r\end{aligned}$$

where $\xi_O \geq \left\|\boldsymbol{\Xi}\left(\mathcal{O}\right)^k\right\|$ for any $k \geq 0$. Note $\xi_O < \infty$ because we can show the existing of the limit of $\{\left\|\boldsymbol{\Xi}\left(\mathcal{O}\right)^0\right\|, \left\|\boldsymbol{\Xi}\left(\mathcal{O}\right)^1\right\|, \ldots\}$ by similar steps in Appendix A.4. Thus

$$\begin{aligned}\mathbb{E}\left[\frac{1}{|\mathcal{X}|^r}\left|\sum_{z_{1:r}\in\mathcal{X}^r}\left(\hat{g}(z_{1:r})-g(z_{1:r})\right)\right|\right] &\leq \mathbb{E}\left[\mathbb{E}\left[\left|\hat{g}(x'_{1:r})-g(x'_{1:r})\right|\left|\mathcal{X}\right.\right]\right]\\ &= \mathbb{E}\left[\left|\hat{g}(x'_{1:r})-g(x'_{1:r})\right|\right]\\ &= \mathbb{E}\left[1_{x'_{1:r}\in\mathcal{K}_\epsilon(\hat{g})}\right]\mathbb{E}\left[\left|\hat{g}(x'_{1:r})-g(x'_{1:r})\right|\left|x'_{1:r}\in\mathcal{K}_\epsilon(\hat{g})\right.\right]\\ &\quad+\mathbb{E}\left[1_{x'_{1:r}\notin\mathcal{K}_\epsilon(\hat{g})}\right]\mathbb{E}\left[\left|\hat{g}(x'_{1:r})-g(x'_{1:r})\right|\left|x'_{1:r}\notin\mathcal{K}_\epsilon(\hat{g})\right.\right]\\ &\leq \xi_\mu\epsilon\cdot 2\xi_g+\epsilon=\left(2\xi_g\xi_\mu+1\right)\epsilon\end{aligned}$$

where $\xi_\mu = \left(\|\boldsymbol{\omega}\|\,\|\boldsymbol{\sigma}\|\,\xi_O\bar{\xi}\right)^r/\underline{\xi}$. By the Markov's inequality, we have

$$\Pr\left[\frac{1}{|\mathcal{X}|^r}\left|\sum_{z_{1:r}\in\mathcal{X}^r}\left(\hat{g}(z_{1:r})-g(z_{1:r})\right)\right|\geq\sqrt{\epsilon}\right]\leq\left(2\xi_g\xi_\mu+1\right)\sqrt{\epsilon} \tag{A.17}$$

Combining (A.16) and (A.17) leads to

$$\begin{aligned}\Pr\left(\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[g]\right|\geq\xi_0\xi^r\sqrt{\epsilon}\right) &\leq \Pr\left(\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[g]\right|\geq\frac{\xi_0\xi^r}{|\mathcal{X}|^r}\left|\sum_{z_{1:r}\in X^r}\left(\hat{g}(z_{1:r})-g(z_{1:r})\right)\right|\right)\\ &\quad+\Pr\left(\frac{1}{|\mathcal{X}|^r}\left|\sum_{z_{1:r}\in\mathcal{X}^r}\left(\hat{g}(z_{1:r})-g(z_{1:r})\right)\right|\geq\sqrt{\epsilon}\right)\\ &\leq \epsilon+\left(2\xi_g\xi_\mu+1\right)\sqrt{\epsilon}\end{aligned}$$

for all $I > I_1$.

From all the above, we have

$$\begin{aligned}&\Pr\left(\left|\mathbb{E}_\infty[g]-\mathbb{E}_{\hat{\mathcal{M}}}[g]\right|\leq 2(\xi_g+1)\epsilon+\xi_0\xi^r\sqrt{\epsilon}\right)\\ &\quad\geq\Pr\left(\left|\mathbb{E}_\infty[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right|\leq\epsilon,\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[g]\right|\leq\xi_0\xi^r\sqrt{\epsilon}\right)\\ &\quad\geq 1-\Pr\left(\left|\mathbb{E}_\infty[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]\right|>\epsilon\right)-\Pr\left(\left|\mathbb{E}_{\hat{\mathcal{M}}}[\hat{g}]-\mathbb{E}_{\hat{\mathcal{M}}}[g]\right|>\xi_0\xi^r\sqrt{\epsilon}\right)\\ &\quad\geq 1-2\epsilon-\left(2\xi_g\xi_\mu+1\right)\sqrt{\epsilon}\end{aligned}$$

for all $I > \max\{I_0, I_1\}$, which yields $\mathbb{E}_{\hat{\mathcal{M}}}[g] \xrightarrow{p} \mathbb{E}_\infty[g]$ due to the arbitrariness of ϵ.

B Settings in applications

B.1 Models

The diffusion processes in Section 6 are driven by the Brownian dynamics

$$\mathrm{d}x_t = -\nabla V(x_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t$$

with $\beta = 0.3$, sample interval 0.002s,

$$V\left(x\right) = \frac{\sum_{i=1}^5\left(|x-c_i|+0.001\right)^{-2}u_i}{\sum_{i=1}^5\left(|x-c_i|+0.001\right)^{-2}}$$

for the one-dimensional process, and $\beta = 2$, sample interval 0.01s,

$$V\left(x\right) = -\log\left(\sum_{i=1}^3 p_i\mathcal{N}\left(x|\mu_i,\Sigma_i\right)\right)$$

for the two-dimensional process, where $c_{1:5} = (-0.3, 0.5, 1, 1.5, 2.3)$, $u_{1:5} = (21, 4, 8, -1, 20)$, $p_{1:3} = (0.25, 0.25, 0.5)$, $\mu_1 = (0, -0.5)$, $\mu_2 = (-1, 0.5)$, $\mu_3 = (1, -0.5)$. The simulation details of alanine dipeptide is given in [3].

B.2 Algorithms

The parameters of discrete OOMs are chosen as: $L = 3$, $m = 10$, $\mathbf{F}_1, \mathbf{F}_2$ are given by the truncated SVD and $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ are indicator functions of all $\mathcal{O}^L$ observation subsequences with length L.

The parameters of binless OOMs are almost the same as discrete ones, except $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ are Gaussian activation functions with random weights of functional link neural networks with $D_1 = D_2 = 100$.

The number of hidden states of HMMs is 10. For continuous data, we partition the state space into 100 discrete bins k-mean clustering (except for the one-dimensional process), and then learn HMMs by the EM algorithm under the assumption that samples within the same bin are drawn independently.

References

[1] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, vol. 4, pp. 2111–2245, 1994.

[2] K. Hornik, M. Stinchcombe, and H.White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[3] B. Trendelkamp-Schroer and F. Noé, "Efficient estimation of rare-event kinetics," *Phys. Rev. X*, vol. 6, pp. 011009, 2016.