GENERATIVE SEQUENCE MODELING FOR INFINITE MEMORY PROCESSES VIA PREDICTIVE STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider estimating the one-step-ahead conditional distribution of a multivariate stochastic process. Many existing approaches rely on assumptions such as finite-range memory, sparsity, or additivity, which can be poorly suited to processes with long-range nonlinear interactions. However, without such structural assumptions, nonparametric estimation is challenging due to the curse of dimensionality. To address this challenge, we introduce a new estimation approach based on the *predictive states* of a process, possibly with infinite-range memory. We show that our estimator achieves fast convergence rates when the past history can be compressed into a low-dimensional statistic that is sufficient for predicting the future. Specifically, we show that the statistical complexity of the estimation problem is determined by the intrinsic dimension of the predictive state space. We establish guarantees for an instantiation of our method based on deep neural network estimators, and we support these theoretical results with experiments.

1 INTRODUCTION

Let X_t, $t \in \mathbb{Z}$, be a stationary process. We aim to learn the one-step-ahead conditional distribution

$$\mathcal{L}(X_{t+1} \mid \mathcal{F}_t), \quad \mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots), \quad t \in \mathbb{Z}. \tag{1}$$

This is a fundamental problem in multi-step forecasting, optimization, and control, because, as we explain below, success at this task lets us learn the predictive distribution for the *whole* future,

$$\mathcal{L}(X_{t+1:\infty} \mid \mathcal{F}_t), \quad t \in \mathbb{Z}.$$

We consider the class of stochastic processes that admit finite-dimensional *predictive states*, which formalize the notion that the information up to time t, $\mathcal{F}_t$, can be compressed into a finite-dimensional statistic, S_t, that is sufficient for predicting the future values of a process. In this sense, finite-dimensional predictive states can be viewed as a dynamic version of sufficient dimension reduction (Li, 1991; Cook, 2007) for stochastic processes.

Our main goal is to develop a recursive estimator of the one-step-ahead conditional distribution that can achieve fast convergence rates when the predictive state space is low-dimensional. This property of a "low-dimensional predictive state space" offers an alternative to other conditions used to mitigate the curse of dimensionality, such as additivity (Stone, 1985), sparsity (Tibshirani, 1996), low intrinsic dimensionality (Bickel & Li, 2007), and hierarchical compositionality (Schmidt-Hieber, 2020). For a discussion of how our work relates to prior work on predictive states, see Appendix A.

1.1 OUR CONTRIBUTIONS

The contributions of this paper are:

1. In Theorem 1, we show that predictive states S_t satisfy $X_{t+1:\infty} \perp\!\!\!\perp X_{-\infty:t} \mid S_t$ for all $t \in \mathbb{Z}$. Also, in Theorem 7 in the Appendix, we show that any predictor R_t which is one-step-ahead predictively sufficient, $X_{t+1} \perp\!\!\!\perp X_{-\infty:t} \mid R_t$, and which can be updated recursively, $R_{t+1} = \tau(R_t, X_{t+1})$ for some (measurable) function τ, is predictively sufficient for the whole future. This generalizes Corollary 2 of Theorem 1 in Shalizi & Crutchfield (2001) from discrete measurable spaces to standard Borel spaces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. In Theorem 2, we show that p-dimensional predictive states S_t can be identified from the conditional expectations of $2p + 1$ random Fourier features (Rahimi & Recht, 2007) of a finite block of future observations. In Theorem 3, we show that these conditional expectations also identify the one-step-ahead conditional distribution of the process.
3. In Theorems 4 and 5, we derive the convergence rate of our estimator in the context of a recurrent neural network (RNN) and mixture density network (MDN). We prove that it can achieve fast convergence rates when the predictive state space is low-dimensional. Additionally, we present results from synthetic experiments to support our theoretical findings.
4. In Theorem 6 in the Appendix, we extend the guarantees of the MDN estimator of Bishop (1994); Zhou et al. (2023) to covariates with possibly low intrinsic dimensionality in the setting of β-mixing processes, which can be of independent interest.

2 Methods and Main Ideas

2.1 Probabilistic Sequence Model

We observe an $\mathcal{X}$-valued sequence X_t, $t = 1, \ldots, n$, for some sample size $n \in \mathbb{N}$, where the unknown observation space $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$ for some fixed $d \in \mathbb{N}$. We consider the model

$$X_{t+1} = G(\varepsilon_{t+1}, S_t), \quad S_t = \tau(S_{t-1}, X_t) = R(X_{-\infty:t}), \quad t \in \mathbb{Z}, \tag{2}$$

where ε_t, $t \in \mathbb{Z}$, is an iid sequence of latent noise inputs satisfying $\varepsilon_{t+1} \perp\!\!\!\perp \mathcal{F}_t$ for all $t \in \mathbb{Z}$, S_t, $t \in \mathbb{Z}$, is an $\mathcal{S}$-valued sequence of latent predictive states, where the unknown predictive state space $\mathcal{S}$ is a Borel subset of $\mathbb{R}^r$ for some ambient dimension $r \in \mathbb{N}$. G is a generation function, τ is a transition function, and R is a reduction function, all of which are unknown and Borel measurable.

Remark 1. *The model from equation 2 includes many nonlinear processes with infinite-range memory, as well as finite-order Markov processes. It is straightforward to extend this model to include exogenous covariates, but we leave this extension for future work.*

The conditional distribution of X_{t+1} given S_t is given by

$$\mathcal{L}(X_{t+1} \mid S_t) = \mathcal{L}(X_{t+1} \mid \mathcal{F}_t), \quad t \in \mathbb{Z}, \tag{3}$$

almost surely, because $S_t = R(X_{-\infty:t})$ is $\mathcal{F}_t$-measurable and because $\varepsilon_{t+1} \perp\!\!\!\perp \mathcal{F}_t$. By Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)), equation 3 is equivalent to the one-step-ahead predictive sufficiency property,

$$X_{t+1} \perp\!\!\!\perp X_{-\infty:t} \mid S_t, \quad t \in \mathbb{Z}. \tag{4}$$

In fact, S_t has the strongest possible form of predictive sufficiency, as shown by Theorem 1. We give a more general result in Theorem 7 in the Appendix, where $\mathcal{X}, \mathcal{S}$ can be standard Borel spaces.

Theorem 1. *The predictive states from equation 2 have the strong predictive sufficiency property,*

$$X_{t+1:\infty} \perp\!\!\!\perp X_{-\infty:t} \mid S_t, \quad t \in \mathbb{Z}. \tag{5}$$

The *prediction process* (Knight, 1975; Aldous, 1981; Löhr, 2010),

$$\Pi_t = \mathcal{L}(X_{t+1:\infty} \mid \mathcal{F}_t), \quad t \in \mathbb{Z}, \tag{6}$$

characterizes the future values of a stochastic process given its past. By Theorem 1 and the aforementioned Doob's conditional independence property,

$$\Pi_t = \mathcal{L}(X_{t+1:\infty} \mid S_t), \quad t \in \mathbb{Z}, \tag{7}$$

almost surely, so by the chain rule, predictive sufficiency, and the recursive update in equation 2,

$$\Pi_{t,L} = \mathcal{L}(X_{t+1:t+L} \mid \mathcal{F}_t) = \mathcal{L}(X_{t+1:t+L} \mid S_t) = \prod_{j=1}^{L} \mathcal{L}(X_{t+j} \mid S_{t+j-1}), \quad t \in \mathbb{Z}, \tag{8}$$

almost surely, for arbitrary horizon $L \in \mathbb{N}$, where the product denotes iterated composition of conditional probability kernels. Hence, for multi-step generative sequence modeling, it suffices to characterize the one-step-ahead conditional distributions $\mathcal{L}(X_{t+1} \mid S_t)$; see Section 2.3.

Unfortunately, we do not observe S_t. We therefore aim to find an estimable, injective map Φ that sends the predictive state values $s \in \mathcal{S}$ to elements in some feature space $\Phi(s) \in \Phi(\mathcal{S})$. As we will show, this estimate of $\Phi(S_t)$ can be used to estimate the conditional distribution of X_{t+1} given S_t.

2.2 Dynamic Embeddings of Predictive States

For some $m \in \mathbb{N}_0$, denote the conditional joint distribution of the next $m+1$ observations by

$$P_{S_t} = \mathcal{L}(X_{t+1:t+m+1} \mid \mathcal{F}_t) = \mathcal{L}(X_{t+1:t+m+1} \mid S_t), \quad t \in \mathbb{Z}. \tag{9}$$

The family of distributions $\{P_s : s \in \mathcal{S}\}$ must satisfy some identifiability conditions, so that every value $s \in \mathcal{S}$ corresponds to a unique distribution P_s. Specifically, we need m to be large enough that every predictive state leads to a distinct $(m+1)$-step-ahead distribution.

We now present the ingredients used to injectively map $\mathcal{S}$ into $[-1,1]^k$ for some $k \geq 2p+1$. We randomly draw $k \geq 2p+1$ real-valued test functions $\varphi_1, \ldots, \varphi_k$ independently of the data from a distribution over a function class. Let $x_1, \ldots, x_{m+1} \in \mathcal{X}$ and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathcal{X}^{m+1}$. In particular, we use random Fourier features (Rahimi & Recht, 2007) of the form

$$\varphi_i(x) = \cos\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j + \alpha_i\right), \tag{10}$$

where $\Omega_{i,j} \overset{\text{iid}}{\sim} N(0, I_d)$ and $\alpha_i \overset{\text{iid}}{\sim} U(-\pi, \pi)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m+1$. Denote $\varphi = (\varphi_1, \ldots, \varphi_k) : \mathcal{X}^{m+1} \to [-1,1]^k$ and let

$$Y_{t+1} = \varphi(X_{t+1:t+m+1}), \quad t \in \mathbb{Z}. \tag{11}$$

Consider the mapping $\Phi : \mathcal{S} \to [-1,1]^k$ defined as

$$\Phi(s) = \int_{\mathcal{X}^{m+1}} \varphi(x) dP_s(x). \tag{12}$$

For each $t \in \mathbb{Z}$, the predictive states S_t are mapped into feature space $\Phi(S_t)$ via the quantity

$$Z_t = \Phi(S_t) = \int_{\mathcal{X}^{m+1}} \varphi(x) d\mathcal{L}(X_{t+1:t+m+1} \mid \mathcal{F}_t)(x) = \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t], \tag{13}$$

where the first and third equalities are by definition, and the second is by equation 12 and Theorem 1.

To guarantee that the mapping Φ is (almost surely) one-to-one, we must limit the geometric complexity of the predictive state space $\mathcal{S}$. In the simplest case, we can take $\mathcal{S}$ to be a Cartesian product of closed, bounded intervals $\mathcal{S} = \prod_{j=1}^{r}[a_j, b_j] \subset \mathbb{R}^r$ for real numbers $a_j < b_j$, $j = 1, \ldots, r \equiv p$. More generally, we only need to be able to characterize $\mathcal{S}$ by a finite number of polynomial equalities and inequalities. A more precise statement uses the notion of a "semialgebraic set"; we simplify the definition from Bochnak et al. (1998) to Euclidean spaces.

Definition 1 (Definition 2.1.4 in Bochnak et al. (1998))**.** *Let $r, a, b_1, \ldots, b_a \in \mathbb{N}$. A semialgebraic subset of $\mathbb{R}^r$ is a subset of the form $\bigcup_{i=1}^{a} \bigcap_{j=1}^{b_i} \{x \in \mathbb{R}^r \mid f_{i,j}(x) *_{i,j} 0\}$, where $f_{i,j} \in \mathbb{R}[x_1, \ldots, x_r]$, $*_{i,j}$ is either $<$ or $=$, for $i = 1, \ldots, a$ and $j = 1, \ldots, b_i$, and $\mathbb{R}[x_1, \ldots, x_r]$ is the ring of polynomials in the variables $x_1, \ldots, x_r$ with coefficients in $\mathbb{R}$.*

We now have the materials needed to state our first assumption. Denote by $\psi_s(\omega)$ the characteristic function of P_s evaluated at $\omega \in \mathbb{R}^{(m+1)\times d}$. The second condition is that $m \in \mathbb{N}_0$ from equation 9 is large enough to identify each $s \in \mathcal{S}$, so for two different predictive state values $s \neq s'$ we have $P_s \neq P_{s'}$. Note that we require the sample size to satisfy $n > 2m+1$.

Assumption 1. *The following conditions hold:*

1. *$\mathcal{S}$ is a compact, semialgebraic subset of $\mathbb{R}^r$ with Hausdorff dimension $\dim_H(\mathcal{S}) = p$.*
2. *The map $s \mapsto P_s$ is one-to-one.*
3. *The map $(s, \omega) \mapsto \psi_s(\omega)$ admits a jointly real analytic extension to $O \times \mathbb{R}^{(m+1)\times d}$, where $O \subset \mathbb{R}^r$ is an open neighborhood of $\mathcal{S}$.*

The following result establishes that Φ is one-to-one almost surely. The key ingredient is the finite witness theorem from Amir et al. (2023). Note that $k \geq 2p+1$, not necessarily $k = 2p+1$.

Theorem 2. *Suppose Assumption 1 holds. Then, with probability one over the draw of $k \geq 2p+1$ random Fourier features $\varphi = (\varphi_1, \ldots, \varphi_k)$, the mapping $s \mapsto \Phi(s)$ is one-to-one.*

Going forward, we condition on a fixed realization of the random Fourier features. We make the following weak assumption, which holds with probability one by Theorem 2.

Assumption 2. *Assume that the mapping $s \mapsto \Phi(s)$ is one-to-one.*

The following result establishes the equality of the conditional distributions given S_t and Z_t.

Theorem 3. *Suppose Assumptions 1 and 2 hold. Then, almost surely,*

$$Q_{Z_t} = \mathcal{L}(X_{t+1} \mid Z_t) = \mathcal{L}(X_{t+1} \mid S_t), \quad t \in \mathbb{Z}.$$

If $(Z_t)_t$ from equation 13 can be estimated well, then we can estimate $Q_z = \mathcal{L}(X_{t+1} \mid Z_t = z)$ from Theorem 3 with a conditional distribution estimator using the estimates $(\hat{Z}_t)_t$ of $(Z_t)_t$.

2.3 Generative Sequence Modeling

Algorithm 1 presents our two-stage estimation method. In practice, the predictive state space's dimension $\dim_H(\mathcal{S}) = p$ is unknown, so we treat the number k of random Fourier features φ as a hyperparameter to be selected from some candidate set $\mathcal{K} \subset \mathbb{N}$, $\min(\mathcal{K}) \geq 3$. The value of $k \in \mathcal{K}$ and the estimator hyperparameters are then selected jointly by optimizing some criterion. We are, here, agnostic about *which* criterion, though a form of cross-validated log-likelihood suggests itself.

Algorithm 1 Learning the one-step-ahead conditional distribution via predictive states

1: **Input:** Observations $(X_t)_{t=1}^n$, random Fourier feature generator, candidate set $\mathcal{K}$, conditional expectation estimator, and conditional distribution estimator
2: Select value of $k \in \mathcal{K}$ and estimator hyperparameters by optimizing a criterion
3: $(\hat{Z}_t)_t \leftarrow$ Fit and apply conditional expectation estimator of $(Z_t)_t$ from equation 13
4: $\hat{Q}_z \leftarrow$ Fit and apply conditional distribution estimator of Q_z from Theorem 3 using $(\hat{Z}_t, X_{t+1})_t$
5: **Output:** Fitted estimators and estimates $\hat{Q}_z$, $(\hat{Z}_t)_t$

In Appendix A, we give an algorithm for estimating functionals of the conditional distribution from equation 8 for arbitrary horizons, e.g., conditional quantiles in multi-step forecasting, conditional risk functionals in optimization, and conditional utilities of action sequences in control.

3 Theory

In this section, we establish the rates of convergence for an instantiation of our method based on the recurrent neural network (RNN) from Chen et al. (2026) and the mixture density network (MDN) from Zhou et al. (2023). We consider an RNN and MDN for the sake of concreteness, and because they enable very fast sequential sampling. In general, other estimators can be used with our method.

3.1 Recurrent Neural Network Estimator

Define $\tau^{(m+1)}$ as the recursive application of τ from equation 2 a total of $(m+1)$ times, so that

$$S_{t+m+1} = \tau^{(m+1)}(S_t, X_{t+1:t+m+1}) = \tau(\cdots \tau(\tau(S_t, X_{t+1}), X_{t+2}) \cdots, X_{t+m+1}). \tag{14}$$

The following result is used to establish the convergence rate guarantees of the RNN estimator.

Lemma 1. *Suppose Assumptions 1 and 2 hold, and that the predictive states S_t, $t \in \mathbb{Z}$, satisfy the recursive update rule from equation 2. Then, Y_{t+1} from equation 11 admits the representation*

$$Y_{t+1} = g^*(Y_{t-m}, \xi_{t-m}, X_{t-m:t}) + \xi_{t+1}, \quad t \in \mathbb{Z},$$

where the noise process ξ_{t+1} is from equation 16 and the function $g^*(\cdot) : \mathcal{D} \to \Phi(\mathcal{S})$ with domain

$$\mathcal{D} = \{(y, e, x) \in [-1,1]^k \times [-2,2]^k \times \mathcal{X}^{m+1} : y - e \in \Phi(\mathcal{S}), \tau^{(m+1)}(\Phi^{-1}(y-e), x) \in \mathcal{S}\},$$

where $\tau^{(m+1)}$ is from equation 14, is defined as

$$g^*(y, e, x) = \Phi(\tau^{(m+1)}(\Phi^{-1}(y - e), x)). \tag{15}$$

We require the following compactness condition on the unknown set $\mathcal{X}$ from equation 2.

Assumption 3. *Assume $X_t \in \mathcal{X}$ almost surely for every $t \in \mathbb{Z}$ for some compact set $\mathcal{X} \subset \mathbb{R}^d$.*

Let $\mathcal{X}_j$ be the j-th dimension of $\mathcal{X}$ for each $j \in [d]$, and denote $\mathcal{X} = \mathcal{X}_{1:d}$. For the noise process

$$\xi_{t+1} = Y_{t+1} - Z_t, \quad t \in \mathbb{Z}, \tag{16}$$

which satisfies $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0$, the features Y_{t+1} from equation 11 can be written as

$$Y_{t+1} = Z_t + \xi_{t+1}, \quad t \in \mathbb{Z}. \tag{17}$$

Note that $Y_t \in [-1, 1]^k$ and $Z_t \in \mathcal{Z} = \Phi(\mathcal{S}) \subset [-1, 1]^k$, so $\xi_t \in [-2, 2]^k$ by equation 16.

Next, we need a smoothness condition for $g^*(\cdot)$ from equation 15 analogous to Assumption 2 in Chen et al. (2026), which involves the definition of a Hölder ball of functions.

Definition 2. *Let $\eta, c > 0$ and let A be a subset of some Euclidean space. Define $\mathcal{H}^\eta(A, c)$ as*

$$\left\{ f : A \to \mathbb{R} \;\middle|\; \max_{\alpha:\|\alpha\|_1 \le \tilde{\eta}} \sup_{x \in A} |D^\alpha f(x)| + \max_{\alpha:\|\alpha\|_1 = \tilde{\eta}} \sup_{\substack{x, x' \in A \\ x \ne x'}} \frac{|D^\alpha f(x) - D^\alpha f(x')|}{\|x - x'\|_2^{\eta - \tilde{\eta}}} \le c \right\},$$

where $\tilde{\eta}$ is the largest integer less than η. For a vector-valued function $f = (f_1, \ldots, f_k)^\top : A \to \mathbb{R}^k$, we write $f \in (\mathcal{H}^\eta(A, c))^k$ if $f_j \in \mathcal{H}^\eta(A, c)$ for every $j = 1, \ldots, k$.

Assumption 4. *Assume that there exist constants $b_{\mathrm{bnd}} \ge 4$ and $c_{\mathrm{Höl}} > 0$ such that $\mathcal{X} \subseteq [-b_{\mathrm{bnd}}, b_{\mathrm{bnd}}]^d$ and $g^*(\cdot)$ from equation 15 admits an extension $\tilde{g}^*(\cdot)$ to the domain $[-b_{\mathrm{bnd}}, b_{\mathrm{bnd}}]^{2k + d(m+1)}$ that is η-Hölder smooth. That is, there exists some $\eta \in \mathbb{N}$, such that*

$$\tilde{g}^*(\cdot) \in \left(\mathcal{H}^\eta \left([-b_{\mathrm{bnd}}, b_{\mathrm{bnd}}]^{2k + d(m+1)}, c_{\mathrm{Höl}} \right) \right)^k.$$

The following invertibility assumption, analogous to Assumption 4 in Chen et al. (2026), is required.

Assumption 5. *For all $i, j \in [k]$, let $\tilde{g}_i^*(\cdot)$ be the i-th component of $\tilde{g}^*(\cdot)$ and let*

$$\Delta_{i,j} = \operatorname*{ess\,sup}_{(y,e,x) \in [-b_{\mathrm{bnd}}, b_{\mathrm{bnd}}]^{2k + d(m+1)}} \left| \frac{\partial \tilde{g}_i^*(y, e, x)}{\partial e_j} \right|,$$

and assume that $\max_{i \in [k]} \sum_{j=1}^k \Delta_{i,j} < 1$.

Next, we introduce the deep neural network (DNN) function class used in the RNN. Let $\sigma_b(x) = [\max(x_i - b_i, 0)]_i$ be the shifted ReLU activation. For $a, b, \ell, w \in \mathbb{N}$ and $c_{\mathrm{bnd},1}, c_{\mathrm{bnd},2} > 0$, let $\mathcal{M}_{n,a,b}^{\mathrm{DNN}}(\ell, w, c_{\mathrm{bnd},1}, c_{\mathrm{bnd},2})$ denote the class of ReLU DNNs $f : \mathbb{R}^a \to \mathbb{R}^b$ of the form

$$f(x) = A_\ell \sigma_{b_\ell} A_{\ell-1} \sigma_{b_{\ell-1}} \cdots A_1 \sigma_{b_1} A_0 x, \tag{18}$$

where the network has ℓ hidden layers, each hidden layer has width at most w, and

$$\max \left(\max_{0 \le j \le \ell} \|A_j\|_\infty, \max_{1 \le j \le \ell} \|b_j\|_\infty \right) \le c_{\mathrm{bnd},1}, \qquad \|f\|_\infty \le c_{\mathrm{bnd},2}. \tag{19}$$

We now introduce the RNN function class. As in Chen et al. (2026), we use a single recurrent layer.

Assumption 6. *Assume there exists $\epsilon_{\mathrm{RNN}} \in (0, 1)$ such that, for every $n \in \mathbb{N}$, the RNN class is*

$$\mathcal{M}_n^{\mathrm{RNN}} = \left\{ \begin{array}{l} (\rho(\cdot), W_h, v_h, \lambda(\cdot)) : \rho(\cdot) \in \mathcal{M}_{n, d(m+1)+k, w_{h,n}}^{\mathrm{DNN}}(\ell_n, w_n, c_{\mathrm{RNN},1}, c_{\mathrm{RNN},2}), \\ \lambda(\cdot) \in \mathcal{M}_{n, w_{h,n}, k}^{\mathrm{DNN}}(\ell_n, w_n, c_{\mathrm{RNN},1}, 1), \quad W_h \in \mathbb{R}^{w_{h,n} \times w_{h,n}} \text{ is upper triangular}, \\ \|W_h\|_\infty \le c_{\mathrm{RNN},1}, \quad \|\mathrm{diag}(W_h)\|_\infty \le 1 - \epsilon_{\mathrm{RNN}}, \quad \|v_h\|_\infty \le c_{\mathrm{RNN},1}, \end{array} \right\},$$

where $c_{\mathrm{RNN},1} = c_{\mathrm{RNN},2} = n^{5\eta+5}$ with η from Assumption 4.

We obtain the RNN parameter estimates by minimizing the sum of squared prediction errors,

$$(\hat{\rho}(\cdot), \hat{W}_h, \hat{v}_h, \hat{\lambda}(\cdot)) \in \underset{(\rho(\cdot), W_h, v_h, \lambda(\cdot)) \in \mathcal{M}_n^{\mathrm{RNN}}}{\arg\min} \sum_{t=m+1}^{n-m-1} \|Y_{t+1} - \lambda(H_{t+1})\|_2^2, \quad \text{subject to } H_{t+1} = \sigma_{v_h}(W_h H_t + \rho(I_{t+1})) \text{ and } H_{m+1} = 0, \tag{20}$$

where $I_{t+1} = (X_{t-m:t}, Y_{t-m})$ and $\mathcal{M}_n^{\mathrm{RNN}}$ is from Assumption 6. See Appendix A for discussion.

3.2 Mixture Density Network Estimator

Recall $Q_z = \mathcal{L}(X_{t+1} \mid Z_t = z)$, $z \in \mathcal{Z} = \Phi(\mathcal{S})$ from Theorem 3. We assume, for all $z \in \mathcal{Z}$, Q_z is absolutely continuous w.r.t. Lebesgue measure, and a jointly measurable version of its density exists

$$f_{X_{t+1}|Z_t}(x|z), \quad z \in \mathcal{Z}. \tag{21}$$

The following notation will be useful. For each $j = 1, \ldots, d$, define $\tilde{\mathcal{Z}}_j = \mathcal{Z} \times \mathcal{X}_{1:j-1}$ and let

$$\tilde{Z}_{t,j} = (Z_t^\top, X_{t+1,1}, \ldots, X_{t+1,j-1})^\top, \quad \tilde{z}_j = (z^\top, x_1, \ldots, x_{j-1})^\top \in \tilde{\mathcal{Z}}_j, \tag{22}$$

where we define $\tilde{Z}_{t,1} = Z_t$ and $\tilde{z}_1 = z \in \mathcal{Z}$. Denote the factorization of $f_{X_{t+1}|Z_t}(x|z)$ by

$$f_{X_{t+1}|Z_t}(x|z) = \prod_{j=1}^{d} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j), \tag{23}$$

where we assume there exist jointly measurable versions of $f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)$, $j = 1, \ldots, d$. Note that, for any Borel set $B \in \mathcal{B}(\mathcal{X})$, we have, almost surely, that

$$Q_{Z_t}(B) = \mathcal{L}(X_{t+1} \mid S_t)(B) = \mathcal{L}(X_{t+1} \mid Z_t)(B) = \int_B f_{X_{t+1}|Z_t}(x|Z_t)dx. \tag{24}$$

To simplify the notation below, all suprema and infima are taken over the supports of the corresponding random variables, and the mixture weights satisfy $\sum_{g=1}^{M_j} \alpha_{g,j}(\tilde{z}_j) = 1$ and $\alpha_{g,j}(\tilde{z}_j) \geq 0$.

The next assumption is analogous to Assumption 2 from Zhou et al. (2023). The first condition is that each conditional density function from equation 23 can be approximated by a conditional Gaussian mixture model. The remaining conditions are standard boundedness and smoothness conditions.

Assumption 7. *Assume that:*

1. *There exist constants $\omega_1, c_{\mathrm{gmix}} > 0$, such that, for every $j \in [d]$ and every $M_j \in \mathbb{N}$, there exist functions $\alpha_{g,j} = \alpha_{g,j}^{(M_j)}$, $\mu_{g,j} = \mu_{g,j}^{(M_j)}$, and $\sigma_{g,j} = \sigma_{g,j}^{(M_j)}$ satisfying*

$$\sup_{x_j, \tilde{z}_j} \left| f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j|\tilde{z}_j) - \sum_{g=1}^{M_j} \frac{\alpha_{g,j}(\tilde{z}_j)}{\sqrt{2\pi}\sigma_{g,j}(\tilde{z}_j)} \exp\left(-\frac{(x_j - \mu_{g,j}(\tilde{z}_j))^2}{2\sigma_{g,j}^2(\tilde{z}_j)} \right) \right| \leq c_{\mathrm{gmix}} M_j^{-\omega_1}.$$

2. *There exists a constant $c_\mu > 0$ such that for all $j \in [d]$,*

$$\max_{g \in [M_j]} \sup_{\tilde{z}_j} |\mu_{g,j}(\tilde{z}_j)| \leq c_\mu \quad \textit{for every } M_j \in \mathbb{N},$$

and there exist constants $c_\sigma^+, c_\sigma^- > 0$ and $\omega_2 > 0$ such that for all $j \in [d]$,

$$c_\sigma^- M_j^{-\omega_2} \leq \min_{g \in [M_j]} \inf_{\tilde{z}_j} \sigma_{g,j}(\tilde{z}_j) \leq \max_{g \in [M_j]} \sup_{\tilde{z}_j} \sigma_{g,j}(\tilde{z}_j) \leq c_\sigma^+ \quad \textit{for every } M_j \in \mathbb{N}.$$

3. *There exist $\gamma \in \mathbb{N}$, $c_{\mathrm{Sob}} > 0$ and $b_{\mathcal{X}} \geq 1$, such that $\mathcal{X} \subseteq [-b_{\mathcal{X}}, b_{\mathcal{X}}]^d$ and, for all $j \in [d]$, $M_j \in \mathbb{N}$, and $g \in [M_j]$, the functions $\alpha_{g,j}(\cdot)$, $\mu_{g,j}(\cdot)$, and $\sigma_{g,j}(\cdot)$ admit continuous extensions $\tilde{\alpha}_{g,j}(\cdot)$, $\tilde{\mu}_{g,j}(\cdot)$, and $\tilde{\sigma}_{g,j}(\cdot)$, respectively, to the domain $B_j = [-b_{\mathcal{X}}, b_{\mathcal{X}}]^{k+j-1}$ such that each extension lies in the Sobolev ball*

$$\left\{ h : \max_{\nu, \|\nu\|_1 \leq \gamma} \operatorname*{ess\,sup}_{\tilde{z}_j \in B_j} |D^\nu h(\tilde{z}_j)| \leq c_{\mathrm{Sob}} \right\},$$

where the maximum is taken over all $(k+j-1)$-dimensional $\mathbb{N}_0$-valued vectors ν whose entries sum to at most γ, and $D^\nu h$ denotes the weak derivative (Giné & Nickl, 2015).

4. There exists a constant $c_f^- > 0$ so that, for all $j \in [d]$, $\inf_{x_j,\tilde{z}_j} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j|\tilde{z}_j) \geq c_f^-$.

Note that, by parts 1 and 2 of Assumption 7 with $M_j = 1$, there exists a constant $c_f^+ > 0$ such that

$$\sup_{x_j,\tilde{z}_j} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j|\tilde{z}_j) \leq c_f^+ = c_{\text{gmix}} + 1/(\sqrt{2\pi}c_\sigma^-) < \infty.$$

The next condition is analogous to Assumption 3 in Zhou et al. (2023). Note that we suppress the dependence of $\alpha_{g,j}(\cdot)$, $\mu_{g,j}(\cdot)$, and $\sigma_{g,j}(\cdot)$ on the neural network parameters to simplify the notation.

Assumption 8. *Assume that there exist constants $c_{\text{MDN}} > 0$, $\kappa \in (0, 1]$, and a sequence $(\Lambda_n)_{n\in\mathbb{N}}$, $\Lambda_n \geq 1$, such that, for every $n \in \mathbb{N}$, $j \in [d]$, the MDN function class is*

$$\mathcal{M}_{n,j}^{\text{MDN}} = \left\{ \begin{array}{l} f_j(x_j|\tilde{z}_j) = \sum_{g=1}^{M_j} \frac{\alpha_{g,j}(\tilde{z}_j)}{\sqrt{2\pi}\sigma_{g,j}(\tilde{z}_j)} \exp\left(-\frac{(x_j - \mu_{g,j}(\tilde{z}_j))^2}{2\sigma_{g,j}^2(\tilde{z}_j)}\right) : \\ c_{\text{MDN}}^{-1} \leq \inf_{x_j,\tilde{z}_j} f_j(x_j|\tilde{z}_j) \leq \sup_{x_j,\tilde{z}_j} f_j(x_j|\tilde{z}_j) \leq c_{\text{MDN}}, \\ \sup_{\tilde{z}_j,g} |\mu_{g,j}(\tilde{z}_j)| \leq c_{\text{MDN}},\ c_{\text{MDN}}^{-1} M_j^{-\omega_2} \leq \inf_{\tilde{z}_j,g} \sigma_{g,j}(\tilde{z}_j) \leq \sup_{\tilde{z}_j,g} \sigma_{g,j}(\tilde{z}_j) \leq c_{\text{MDN}}, \\ |\log f_j(x_j \mid z, u_j) - \log f_j(x_j \mid z', u_j)| \leq \Lambda_n \|z - z'\|_2^\kappa \end{array} \right\},$$

where the last condition holds for every $x_j \in \mathcal{X}_j$, $u_j \in \mathcal{X}_{1:j-1}$, and $z, z' \in [-1, 1]^k$. The functions $\mu_{g,j}(\cdot)$ and $\sigma_{g,j}(\cdot)$ are parametrized by DNNs. Similarly, the mixture weights $\alpha_{g,j}(\cdot)$ are obtained by a DNN followed by a normalization $\alpha_{g,j} = a_{g,j} / \sum_{i=1}^{M_j} a_{i,j}$ to ensure $\sum_{g=1}^{M_j} \alpha_{g,j}(\tilde{z}_j) = 1$, where the DNN outputs satisfy $0 \leq a_{i,j}(\tilde{z}_j) \leq c_a$ and $\inf_{\tilde{z}_j} \sum_{i=1}^{M_j} a_{i,j}(\tilde{z}_j) \geq c_\alpha$ for some $c_a, c_\alpha > 0$.

Recall γ from Assumption 7. Assume the number of parameters in the j-th MDN is proportional to

$$M_j^{(k+j-1+\gamma)/\gamma} n^{(k+j-1)/(2\gamma+k+j-1)} \log(M_j n),$$

and that the DNNs for the coefficients use ReLU activations where the number of hidden layers is $O\left(\log(\frac{M_j}{\epsilon_{n,j}}) + 1\right)$ and the numbers of weight parameters and hidden units are each $O\left(M_j^{(k+j-1)/\gamma} \epsilon_{n,j}^{-(k+j-1)/\gamma} (\log(\frac{M_j}{\epsilon_{n,j}}) + 1)\right)$, where $\epsilon_{n,j} = n^{-\gamma/(2\gamma+k+j-1)}$.

The next condition controls the average probability the Gaussian MDN assigns outside of the support $\mathcal{X}_j$. This is required because of the compactness from Assumption 3 and part (4) of Assumption 7.

Assumption 9. *Recall ω_1 from Assumption 7. Assume there exists a constant $c_{\text{tail}} > 0$, such that, for all $j \in [d]$ and $M_j \in \mathbb{N}$, the functions $\alpha_{g,j} = \alpha_{g,j}^{(M_j)}$, $\mu_{g,j} = \mu_{g,j}^{(M_j)}$, and $\sigma_{g,j} = \sigma_{g,j}^{(M_j)}$ satisfy*

$$\int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} \sum_{g=1}^{M_j} \frac{\alpha_{g,j}(\tilde{z}_j)}{\sqrt{2\pi}\sigma_{g,j}(\tilde{z}_j)} \exp\left(-\frac{(x_j - \mu_{g,j}(\tilde{z}_j))^2}{2\sigma_{g,j}^2(\tilde{z}_j)}\right) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \leq c_{\text{tail}} M_j^{-2\omega_1}.$$

Denote the learned predictive state representations, i.e., the estimated counterpart of $\tilde{Z}_{t,j}$, by

$$\hat{\tilde{Z}}_{t,j} = (\hat{Z}_t^\top, X_{t+1,1}, \ldots, X_{t+1,j-1})^\top, \quad j = 1, \ldots, d, \tag{25}$$

with $\hat{\tilde{Z}}_{t,1} = \hat{Z}_t = \hat{\lambda}(\hat{H}_{t+1})$, where $\hat{H}_{t+1}$ is the hidden state estimate and $\hat{\lambda}(\cdot)$ is the output map from equation 20. The MDN estimator is obtained by maximizing the conditional log-likelihood

$$\hat{f}(x|z) = \prod_{j=1}^{d} \hat{f}_j(x_j \mid \tilde{z}_j), \quad \hat{f}_j \in \operatorname*{arg\,max}_{f_j \in \mathcal{M}_{n,j}^{\text{MDN}}} \sum_{t=m+1}^{n-m-1} \log f_j(X_{t+1,j}|\hat{\tilde{Z}}_{t,j}), \quad j = 1, \ldots, d, \tag{26}$$

where the argmax is taken over the MDN function class $\mathcal{M}_{n,j}^{\text{MDN}}$ from Assumption 8. In practice, an optimization procedure is used. In Theorem 6 in the Appendix, we derive the convergence rate.

The following high-level condition is used to establish the consistency of our two-stage estimator.

Assumption 10. *Assume that there exists a constant $c_{\mathrm{aprx}} > 0$, such that, for every $j \in [d]$, there exists an MDN $g_{n,j} \in \mathcal{M}_{n,j}^{\mathrm{MDN}}$ that satisfies*

$$\sup_{x_j, \tilde{z}_j} \left| g_{n,j}(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \right| \leq c_{\mathrm{aprx}} a_{n,j}, \tag{27}$$

$$\int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} g_{n,j}(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \leq c_{\mathrm{aprx}} a_{n,j}^2, \tag{28}$$

where $a_{n,j} = M_j^{-\omega_1} + M_j^{4\omega_2}\epsilon_{n,j}$ and $\epsilon_{n,j} = n^{-\gamma/(2\gamma+k+j-1)}$.

3.3 Stationarity and Temporal Dependence

The definition of a β-mixing process is required.

Definition 3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $\mathcal{B}$, $\mathcal{C}$ be sub-σ-algebras of $\mathcal{A}$. Define*

$$\beta(\mathcal{B}, \mathcal{C}) = \mathbb{E}\left(\sup_{C \in \mathcal{C}} |\mathbb{P}(C) - \mathbb{P}(C \mid \mathcal{B})| \right).$$

For some process V_t, $t \in \mathbb{Z}$, let $\mathcal{F}_{t_1:t_2} = \sigma(V_t : t_1 \leq t \leq t_2)$. Define the β-mixing coefficient by

$$\beta(j) = \sup_{t \in \mathbb{Z}} \beta(\mathcal{F}_{-\infty:t}, \mathcal{F}_{t+j:\infty}), \quad j \geq 1.$$

The process V_t, $t \in \mathbb{Z}$, is said to be β-mixing (or absolutely regular) if $\beta(j) \to 0$ as $j \to \infty$.

The mixing condition below is used to establish the consistency of the RNN and MDN estimators.

Assumption 11. *Assume that the process (X_t, Y_t, Z_t, ξ_t), $t \in \mathbb{Z}$, is stationary and exponentially β-mixing, i.e., for some $c_{\mathrm{mix}}^1, c_{\mathrm{mix}}^2 > 0$, we have $\beta(j) \leq c_{\mathrm{mix}}^1 \exp(-c_{\mathrm{mix}}^2 j)$, $j \geq 1$.*

Remark 2. *Sufficient conditions for the existence of a stationary distribution for the model from equation 2 have been studied by Doukhan et al. (2023) with a coupling approach, and by Chen & Wu (2016) using the physical dependence measure of Wu (2005).*

3.4 Main Theoretical Result

We now present the main results for the estimator $\hat{f}$ from equation 26.

Theorem 4. *Suppose Assumptions 1–11 hold and the number of random Fourier features satisfies $k \geq 2p+1$. Suppose the number of mixture components $M_j = M$ for every $j \in [d]$. In the RNN estimator from equation 20, take the maximum width of the hidden layers in $\rho(\cdot)$ and $\lambda(\cdot)$ to satisfy*

$$w_n \asymp n^{\max\left(\frac{2k+d(m+1)}{4(2k+d(m+1))+2\eta}, \frac{2k+d(m+1)}{5(2k+d(m+1))+3}\right)},$$

the recurrent width $w_{h,n} = \dim(H_t)$ to satisfy $w_{h,n} \asymp w_n \log n$, and the feedforward depth to satisfy $2 \leq \ell_n \leq C \log n$ for a constant $C > 0$. Then, the estimator $\hat{f}$ from equation 26 satisfies

$$\begin{aligned} \left\| \hat{f} - f_{X_{t+1}|Z_t} \right\|_{L^2} &= \left(\int_{\mathcal{Z}} \int_{\mathcal{X}} \left| \hat{f}(x \mid z) - f_{X_{t+1}|Z_t}(x \mid z) \right|^2 dx dP_{Z_t}(z) \right)^{1/2} \\ &= O_{\mathbb{P}}\begin{pmatrix} M^{-\omega_1} + M^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM) \\ + \Lambda_n^{1/2} n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/4} \end{pmatrix}. \end{aligned} \tag{29}$$

If Λ_n from Assumption 8 is $O(1)$, and the number of mixture components is an integer that satisfies

$$M \asymp n^{\frac{2\gamma^2}{(2\gamma+k+d-1)((2\omega_1+1+8\omega_2)\gamma+k+d-1)}},$$

then the estimator $\hat{f}$ from equation 26 has the convergence rate

$$\left\| \hat{f} - f_{X_{t+1}|Z_t} \right\|_{L^2} = O_{\mathbb{P}}\begin{pmatrix} n^{-\frac{2\gamma^2\omega_1}{(2\gamma+k+d-1)((2\omega_1+1+8\omega_2)\gamma+k+d-1)}} \log^3 n \\ + n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/4} \end{pmatrix}.$$

Remark 3. *Theorem 4 is a statement about $\hat{f}$ evaluated at the true predictive state embedding, while in practice $\hat{f}$ is evaluated at the learned predictive state representation.*

In view of Remark 3, we make a high-level RNN assumption to bound the out-of-sample error.

Theorem 5. *Suppose the network choices and assumptions from Theorem 4 hold, and that the rate from equation 29 tends to zero. Fit the RNN and MDN estimators as in equation 20 and equation 26, respectively, using $X_{1:n}$. For any fixed horizon $L \in \mathbb{N}$, let the prediction $\hat{Z}_{n+L-1} = \hat{\lambda}(\hat{H}_{n+L})$ be obtained by continuing the fitted recurrence with observed inputs through step $n+L-1$, without refitting the RNN. Suppose, for some $C > 0$ and all sufficiently large n, that the RNN satisfies*

$$\mathbb{E}\left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^2 \le C n^{-\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11}. \tag{30}$$

Then, as the training sample size $n \to \infty$,

$$\begin{aligned} &\left(\int_{\mathcal{X}} \left|\hat{f}(x \mid \hat{Z}_{n+L-1}) - f_{X_{n+L}|Z_{n+L-1}}(x \mid Z_{n+L-1})\right|^2 dx\right)^{1/2} \\ &= O_{\mathbb{P}}\left(\Lambda_n^{\frac{k}{2\kappa+k}} \left[\begin{array}{l} M^{-\omega_1} + M^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM) \\ + \Lambda_n^{1/2} n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/4} \end{array}\right]^{\frac{2\kappa}{2\kappa+k}}\right). \end{aligned} \tag{31}$$

In particular, this error converges to zero in probability if Λ_n from Assumption 8 is $O(1)$.

Remark 4. *We suspect equation 30 can be proven by extending Lemma 11 in the supplement of Chen et al. (2026) under more assumptions, but due to space limitations this is left for future work.*

4 Experiments

We implement Algorithm 1 in PyTorch. As a baseline, we compare with an MDN whose lag window is a hyperparameter, where the largest possible lag grows with n. This is commonly done to accommodate processes with infinite-range memory. We consider the data generating processes (DGPs) below. The matrices $A_1^S, A_2^S \in \mathbb{R}^{r\times r}$, $A_1^X, A_2^X \in \mathbb{R}^{r\times d}$ are chosen so each process has a unique stationary distribution. We use truncated normal conditional distributions with $B_{\text{trunc}} = [-2.5, 2.5]^d$. The dimensions are $r = p = d = 3$, and $\tanh(\cdot)$, $\cos(\cdot)$ are applied coordinate-wise.

$$\textbf{DGP 1:} \quad X_{t+1} \sim \text{TN}(S_t, I_d; B_{\text{trunc}}), \quad S_t = \tanh(A_1^S S_{t-1} + A_1^X X_t).$$
$$\textbf{DGP 2:} \quad X_{t+1} \sim \text{TN}(S_t, I_d; B_{\text{trunc}}), \quad S_t = \cos(A_2^S S_{t-1} + A_2^X X_t).$$

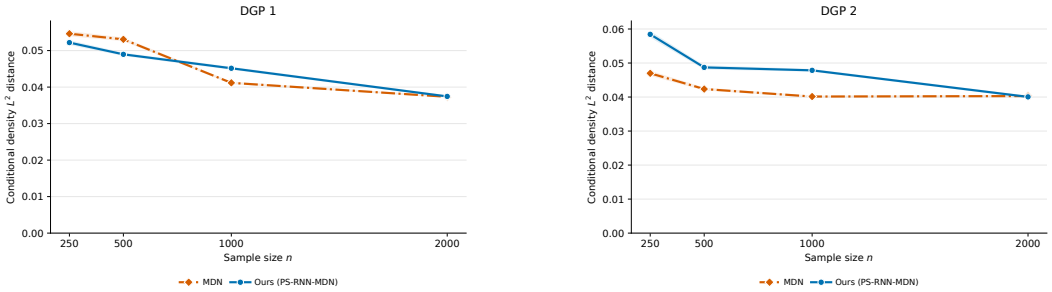


Figure 1: L^2 distance by sample size n. We fit each estimator on the first n samples, we calculate $(L_{\text{eval}}^{-1} \sum_{t=n}^{n+L_{\text{eval}}-1} \int_{B_{\text{trunc}}} (\hat{f}_t(x) - f_t(x))^2 dx)^{1/2}$, where $L_{\text{eval}} = 500$, f_t is the true conditional density, and $\hat{f}_t \in \{\hat{f}_t^{\text{Ours}}, \hat{f}_t^{\text{MDN}}\}$ is the fitted conditional density, then we average this distance over 100 replications. Our method performs comparably to the MDN with a growing lag window.

5 Conclusion and Future Work

We introduce a generative sequence modeling framework based on predictive states. We derive our estimator's convergence rate, and show that fast rates are achievable when the predictive states are low-dimensional. For future work, it would be interesting to apply our ideas to language modeling.

AI USE STATEMENT

Starting from core ideas of the authors, we acknowledge that generative AI was used to: identify and search for relevant literature, brainstorm ideas involving the conceptual frameworks, suggest edits to the paper to improve prose, suggest edits to the proofs to check correctness, shorten length, and improve readability, suggest edits to software code for the method, experiments, and figures, suggest experimental parameters for experiments, identify possible gaps in the paper, provide feedback on the experiments, and generate the TikZ code for Figure 2.

Generative AI was not used to: draft parts of the paper, format references, formulate the presented mathematical claims, provide the critical ingredients for proving mathematical claims, suggest a structure for the paper, assist with translation, interpret results, generate synthetic data sets, or propose a title or keywords for the paper.

Research subtasks not applicable to this work: propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis, formulate questions for surveys or interviews, create artifacts, transcribe recordings of research material.

The authors have reviewed all content suggested by generative AI and produced with the aid of generative AI. For example, the authors have reviewed all literature found by generative AI and checked the correctness of any suggestions made by generative AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors of this paper have read and adhere to the ICLR Code of Ethics. This work focuses on generative sequence modeling, which is used in a wide range of tasks, such as multi-step forecasting, optimization, and control. Our research does not involve human subjects or the collection of sensitive data. The authors declare that there are no conflicts of interest.

REPRODUCIBILITY STATEMENT

Proofs of the results from the main text are in the Appendix. Additional results and their proofs are also in the Appendix. The code will be made publicly available upon acceptance to facilitate reproducibility and will be made available to reviewers during the review process.

REFERENCES

David J. Aldous. Weak convergence and the general theory of processes. Unpublished draft of monograph; Department of Statistics, University of California, Berkeley, CA 94720, 1981. URL `https://www.stat.berkeley.edu/~aldous/Papers/weak-gtp.pdf`.

Tal Amir, Steven Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. *Advances in Neural Information Processing Systems*, 36:42516–42551, 2023.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Henry C.P. Berbee. Random walks with stationary increments and renewal theory. *Mathematical Centre Tracts*, 1979.

Peter J. Bickel and Bo Li. *Complex Datasets and Inverse Problems*, chapter Local polynomial regression on unknown manifolds, pp. 177—-186. Institute of Mathematical Statistics, 2007.

Martin Bilodeau and David Brenner. *Theory of Multivariate Statistics*. Springer New York, New York, NY, 1999.

Christopher M. Bishop. Mixture density networks. Technical Report, 1994.

Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real Algebraic Geometry*. Springer-Verlag, Berlin, 1998.

Byron Boots, Geoffrey Gordon, and Arthur Gretton. Hilbert space embeddings of predictive state representations, 2013. URL https://arxiv.org/abs/1309.6819.

Fabio Boschetti. Mapping the complexity of ecological models. *Ecological Complexity*, 5(1):37–47, 2008.

Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107 – 144, 2005.

Likai Chen and Wei Biao Wu. Stability and asymptotics for autoregressive processes. *Electronic Journal of Statistics*, 10(2):3723–3751, 2016.

Xiao Chen, Yu Chen, Zhouyu Shen, and Dacheng Xiu. Recurrent neural networks for nonlinear time series. Chicago Booth Research Paper, 2026.

Richard W. Clarke, Mervyn P. Freeman, and Nicholas W. Watkins. Application of computational mechanics to the analysis of natural data: An example in geomagnetism. *Physical Review E*, 67, 2003.

R. Dennis Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63: 105–108, 1989.

Paul Doukhan, Michael H. Neumann, and Lionel Truquet. Stationarity and ergodic properties for some observation-driven models in random environments. *Annals of Applied Probability*, 33(6B): 5145–5170, 2023.

Carlton Downey, Ahmed Hefny, Byron Boots, Geoffrey J. Gordon, and Boyue Li. Predictive state recurrent neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

Peter Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986.

Robert Haslinger, Kristina Lisa Klinkner, and Cosma Rohilla Shalizi. The computational structure of spike trains. *Neural Computation*, 22:121–157, 2010.

Ahmed Hefny, Carlton Downey, and Geoffrey J. Gordon. Supervised learning for dynamical system learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Olav Kallenberg. *Foundations of Modern Probability*. Springer Nature, Switzerland, third edition, 2021.

Frank B. Knight. A predictive view of continuous time processes. *The Annals of Probability*, 3(4): 573—-596, 1975.

Frank B. Knight. *Essays on the Prediction Process*, volume 1 of *Institute of Mathematical Statistics Lecture Notes Series*. Institute of Mathematical Statistics, Hayward, California, 1981. URL http://projecteuclid.org/euclid.lnms/1215464503.

Frank B. Knight. *Foundations of the Prediction Process*. Clarendon Press, Oxford, 1992.

John Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 593–600, 2009.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Michael Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. *Advances in neural information processing systems*, 14, 2001.

Wolfgang Löhr. *Models of discrete-time stochastic processes and associated complexity measures*. PhD thesis, Universität Leipzig, Leipzig, Germany, 2010.

Shahar Mendelson. A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*, pp. 1–40, 2003.

Roberto N. Muñoz, Angus Leung, Aidan Zecevik, Felix A. Pollock, Dror Cohen, Bruno van Swinderen, Naotsugu Tsuchiya, and Kavan Modi. General anesthesia reduces complexity and temporal asymmetry of the informational structures derived from neural recordings in Drosophila. *Physical Review Research*, 2, 2020.

Chrystopher L. Nehaniv and Elena Antonova. Simulating and reconstructing neurodynamics with epsilon-automata applied to electroencephalography (eeg) microstate sequences. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–9, 2017.

Dmitry Nerukh, George Karvounis, and Robert C. Glen. Complexity of classical dynamics of molecular systems. I. Methodology. *The Journal of Chemical Physics*, 117(21):9611–9617, 2002.

A.J. Palmer, C.W. Fairall, and W.A. Brewer. Complexity in the atmosphere. *IEEE Transactions on Geoscience and Remote Sensing*, 38(4):2056–2063, 2000.

Joongwoo Brian Park, Jeong Won Lee, Jae-Suk Yang, Hang-Hyun Jo, and Hie-Tae Moon. Complexity analysis of the stock market. *Physica A: Statistical Mechanics and its Applications*, 379 (1):179–187, 2007.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 2021.

Wesley C. Salmon. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, Pittsburgh, 1971. With contributions by Richard C. Jeffrey and James G. Greeno.

Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1984.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879, 2001.

Cosma Rohilla Shalizi and Cristopher Moore. What is a macrostate? from subjective measurements to objective dynamics. *Foundations of Physics*, 55, 2025.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2001.

Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13 (2):689–705, 1985.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Guillaume Valette. On subanalytic geometry, 2025. URL https://arxiv.org/abs/2507.23622.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arun Venkatraman, Nicholas Rhinehart, Wen Sun, Lerrel Pinto, Martial Hebert, Byron Boots, Kris Kitani, and J. Bagnell. Predictive-state decoders: Encoding the future into recurrent networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Michael Wieck-Sosa and Cosma Rohilla Shalizi. Estimating dynamic models by matching random features, 2026a. URL https://arxiv.org/abs/2607.21916.

Michael Wieck-Sosa and Cosma Rohilla Shalizi. Dynamic models with p parameters are identified by $2p + 1$ random features, 2026b. URL https://arxiv.org/abs/2607.16035.

Wei Biao Wu. Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

Yunzhe Zhou, Chengchun Shi, Lexin Li, and Qiwei Yao. Testing for the markov property in time series via deep conditional generative learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1204–1222, 2023.

A Additional Discussions

A.1 Related Work

Predictive states have a long history, with many independent reinventions of (essentially) the same ideas in different formalisms and different jargons (Salmon, 1971; 1984; Knight, 1975; 1981; 1992; Aldous, 1981; Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001; Grassberger, 1986; Jaeger, 2000; Littman et al., 2001; Langford et al., 2009). In computational mechanics (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001), methods for reconstructing (a version of) predictive states have been used in various fields, such as ecology (Boschetti, 2008), chemistry (Nerukh et al., 2002), neuroscience (Haslinger et al., 2010; Nehaniv & Antonova, 2017; Muñoz et al., 2020), finance (Park et al., 2007), and the Earth Sciences (Palmer et al., 2000; Clarke et al., 2003)

Our focus is related to, but distinct from, computational mechanics (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001). Originally, Crutchfield & Young (1989) focused on equivalence classes of histories that induce the same conditional distribution over future events, referred to as the "causal" states. We instead focus on the $\mathcal{S}$-valued process, S_t, $t \in \mathbb{Z}$, which we refer to as the predictive states, to avoid overloading terms and conflating predictive states with causality in the sense of Pearl (2009); Spirtes et al. (2001) (though see Shalizi & Moore (2025)).

Function-learning approaches for predictive state estimation have also been proposed (Langford et al., 2009). Building on these ideas, Boots et al. (2013) and Hefny et al. (2015) proposed learning predictive states by regressing "sufficient features" of future observations on the past observations. For example, the sufficient features for a Gaussian conditional distribution are based on the first two conditional moments. More recently, Downey et al. (2017) and Venkatraman et al. (2017) considered using the hidden states of recurrent neural networks (RNNs) to encode the predictive states.

We adapt predictive states for the purposes of generative sequence modeling, which plays a central role in modern machine learning largely due to the success of the transformer (Vaswani et al., 2017). Ideally, we would like to characterize each $\mathcal{L}(X_{t+1} \mid \mathcal{F}_t)$ by identifying "sufficient features", as in the aforementioned literature. Unfortunately, outside of simple cases, it is challenging to determine

what these sufficient features should be. As a result, this approach of hand-crafting sufficient features does not readily extend to richer families of distributions with nonstandard behavior.

We show that user-chosen "sufficient features" can be replaced by a small number of random features (Rahimi & Recht, 2007). We emphasize that our use of random features is fundamentally different from their conventional use for approximating kernel methods. In contrast, we use a small number of random features as low-dimensional witness functions (or probes), as we will discuss.

A.2 Conditional Functional Estimation

Algorithm 2 explains how to estimate functionals of the conditional distribution from equation 8. Due to space limitations, we will report the theory and experiments in a separate manuscript.

Algorithm 2 Conditional functional estimation via predictive states

Input: Inputs of Algorithm 1, number of simulations N^{sim}, horizon L, and functional $\Gamma(\cdot)$
Fitted estimators and estimates $\hat{Q}_z, (\hat{Z}_t)_t \leftarrow$ Run Algorithm 1
for $j = 1, \ldots, N^{\text{sim}}$ **do**
    Sample $\hat{X}^{(j)}_{n+1} \sim \hat{Q}_{\hat{Z}_n}$
    for $\ell = 1, \ldots, L-1$ **do**
        Predict $\hat{Z}^{(j)}_{n+\ell}$ based on $(X_{1:n}, \hat{X}^{(j)}_{n+1:n+\ell})$ and sample $\hat{X}^{(j)}_{n+\ell+1} \sim \hat{Q}_{\hat{Z}^{(j)}_{n+\ell}}$
    end for
end for
Obtain estimate of conditional functional $\Gamma(\hat{\Pi}_{n,L})$, where $\hat{\Pi}_{n,L} = \frac{1}{N^{\text{sim}}} \sum_{j=1}^{N^{\text{sim}}} \delta_{\hat{X}^{(j)}_{n+1:n+L}}$
Output: Estimate of conditional functional $\Gamma(\hat{\Pi}_{n,L})$

A.3 Discussion of Recurrent Neural Network Architecture

We provide an explanation of how the RNN architecture works. At each step t, the input $I_{t+1} = (X_{t-m:t}, Y_{t-m})$ is transformed by a deep neural network (DNN) $\rho(\cdot)$. This transformed input is then linearly combined with the linearly transformed current hidden state $W_h H_t$, which is passed through a shifted ReLU activation, σ_{v_h}, defined as $\sigma_{v_h}(x) = [\max(x_i - v_{h,i}, 0)]_{1 \le i \le \dim(H_t)}$, to produce the next hidden state H_{t+1}. A second DNN, $\lambda(\cdot)$, maps H_{t+1} to the output O_{t+1}. Figure 2 (based on Figure 1 in Chen et al. (2026)) illustrates the RNN architecture from Section 3.1.

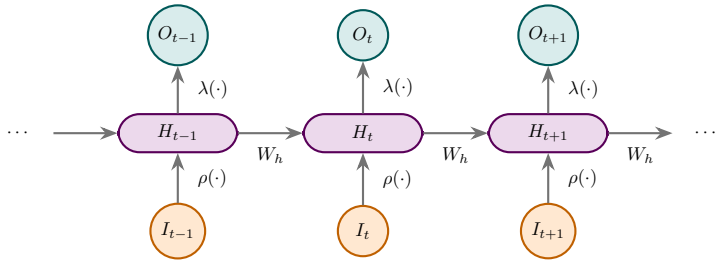


Figure 2: Illustration of the RNN architecture.

B Preliminaries

We state several definitions and results from the literature. We mostly retain the notation from those papers, and we refer to them for proofs.

First, we present the definitions of σ-subanalytic sets and σ-subanalytic functions. These definitions require the definitions of globally subanalytic sets, globally semianalytic sets, and semianalytic sets. Figure 3 (Figure 3 in Amir et al. (2023)) summarizes the relationships among the different classes of sets and functions; see the appendix of Amir et al. (2023) for more discussion.

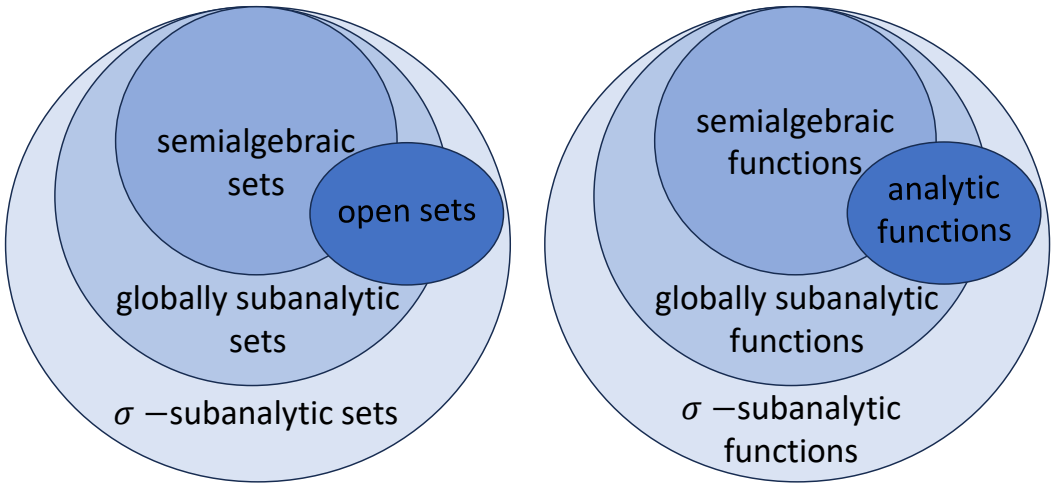


Figure 3: Relationships among classes of sets and functions (Figure 3 in Amir et al. (2023)).

Definition 4 (Definition 1.1.1 in Valette (2025)). *A subset $E \subseteq \mathbb{R}^n$ is called semianalytic if it is locally defined by finitely many real analytic equalities and inequalities. Namely, for each $a \in \mathbb{R}^n$, there is a neighborhood U of a, and real analytic functions f_{ij}, g_{ij} on U, where $i = 1, \dots, r$ and $j = 1, \dots s_i$, such that*

$$E \cap U = \bigcup_{i=1}^{r} \bigcap_{j=1}^{s_i} \{\mathbf{z} \in U | g_{ij}(\mathbf{z}) > 0 \text{ and } f_{ij}(\mathbf{z}) = 0\}. \tag{32}$$

Definition 5 (Definition 1.1.3 in Valette (2025)). *A subset $Z \subseteq \mathbb{R}^n$ is globally semianalytic if $V_n(Z)$ is a semianalytic subset of $\mathbb{R}^n$, where $V_n : \mathbb{R}^n \to (-1, 1)^n$ is the homeomorphism defined by*

$$V_n\left(\mathbf{z} = (z_1, \dots, z_n)\right) = \left(\frac{z_1}{\sqrt{1 + |\mathbf{z}|^2}}, \dots, \frac{z_n}{\sqrt{1 + |\mathbf{z}|^2}}\right).$$

Definition 6 (Definition 1.1.6 in Valette (2025)). *A subset $E \subseteq \mathbb{R}^n$ is globally subanalytic if it can be presented as a linear projection of a globally semianalytic set; more precisely, if there exists a globally semianalytic set $Z \subseteq \mathbb{R}^{n+p}$ such that $E = \pi(Z)$, with $\pi : \mathbb{R}^{n+p} \to \mathbb{R}^n$ being the projection operator that omits the last p coordinates while leaving the remaining coordinates unchanged.*

Definition 7 (Definition A.10 in Amir et al. (2023)). *We say that a subset $A \subseteq \mathbb{R}^D$ is σ-subanalytic if it is a countable union of globally subanalytic subsets of $\mathbb{R}^D$.*

Definition 8 (Definition A.13 in Amir et al. (2023)). *Let $\mathbb{M} \subseteq \mathbb{R}^D$ be a σ-subanalytic set. We say that $f : \mathbb{M} \to \mathbb{R}^L$ is a σ-subanalytic function if its graph is a σ-subanalytic subset of $\mathbb{R}^{D+L}$.*

Second, we state basic properties of σ-subanalytic sets.

Lemma 2 (Proposition A.11 in Amir et al. (2023)). *Assume that $A, B \subseteq \mathbb{R}^D$ and $C \subseteq \mathbb{R}^M$ are σ-subanalytic sets. Then:*

1. *$A \cup B$ is σ-subanalytic. More generally, any countable union of σ-subanalytic sets is σ-subanalytic.*
2. *$A \cap B$ is σ-subanalytic.*
3. *$A \times C$ is σ-subanalytic.*
4. *If $\pi : \mathbb{R}^D \to \mathbb{R}^L$ is a linear projection, then $\pi(A)$ is a σ-subanalytic set.*
5. *A is a* countable *union of C^∞ manifolds.*

Third, we present a corollary of the finite witness theorem (Theorem A.2) from Amir et al. (2023). The notion of dimension that is used is the Hausdorff dimension. The result holds for almost every $(\vartheta_1, \dots, \vartheta_{2D+1})$ with respect to the Lebesgue measure when $\mathbb{W} = \mathbb{R}^q$ or is an open subset of $\mathbb{R}^q$.

Lemma 3 (Corollary A.20 in Amir et al. (2023)). *Let $\mathbb{M} \subseteq \mathbb{R}^r$, $\mathbb{W} \subseteq \mathbb{R}^q$ be σ-subanalytic sets of dimension D and D_ϑ respectively. Let $F : \mathbb{M} \times \mathbb{W} \to \mathbb{R}$ be a σ-subanalytic function. Define the set*

$$\mathcal{N} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{M} \times \mathbb{M} \mid F(\mathbf{x}; \vartheta) = F(\mathbf{y}; \vartheta),\ \forall \vartheta \in \mathbb{W}\}.$$

Suppose that for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{M} \times \mathbb{M} \setminus \mathcal{N}$

$$\dim\{\vartheta \in \mathbb{W} \mid F(\mathbf{x}; \vartheta) = F(\mathbf{y}; \vartheta)\} \leq D_\vartheta - 1. \tag{33}$$

Then for generic $(\vartheta_1, \ldots, \vartheta_{2D+1}) \in \mathbb{W}^{2D+1}$,

$$\mathcal{N} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{M} \times \mathbb{M} \mid F(\mathbf{x}; \vartheta_i) = F(\mathbf{y}; \vartheta_i),\ \forall i = 1, \ldots 2D + 1\}. \tag{34}$$

Moreover, if $\mathbb{W}$ is an open and connected subset of $\mathbb{R}^q$, and $F(\mathbf{x}; \vartheta)$ is analytic as a function of ϑ for all fixed $\mathbf{x} \in \mathbb{M}$, then equation 33 is not required, as it is automatically satisfied.

The next result is the uniqueness theorem for characteristic functions.

Lemma 4 (Proposition 2.9 in Bilodeau & Brenner (1999)). *The characteristic function $c : \mathbb{R}^n \to \mathbb{C}$ of $\mathbf{x}$, defined by $c(\mathbf{t}) = c_{\mathbf{x}}(\mathbf{t}) = \mathbb{E}(e^{i\mathbf{t}'\mathbf{x}})$, is unique:*

$$\mathbf{x} \overset{d}{=} \mathbf{y} \quad \iff \quad c_{\mathbf{x}}(\mathbf{t}) = c_{\mathbf{y}}(\mathbf{t}), \quad \forall \mathbf{t} \in \mathbb{R}^n.$$

The following elementary result is stated as Theorem 4.17 in Rudin (2021).

Lemma 5. *Suppose f is a continuous 1-1 mapping of a compact metric space X onto a metric space Y. Then the inverse mapping f^{-1} defined on Y by*

$$f^{-1}(f(x)) = x, \quad x \in X,$$

is a continuous mapping of Y onto X.

The following result is an approximation bound for deep neural networks (DNNs) stated as Lemma 2 in Zhou et al. (2023). For the following result, let $W^{\gamma,\infty}([-1,1]^d)$ denote the Sobolev space of functions whose weak derivatives up to order γ are bounded, as in Assumption 7.

Lemma 6. *There exists a DNN class $\mathcal{M}^{\mathrm{DNN}}$ with the ReLU activation, such that, for any $0 < \epsilon \leq 1$,*

1. *$\mathcal{M}^{\mathrm{DNN}}$ approximates $W^{\gamma,\infty}([-1,1]^d)$, in the sense that, for any $g^* \in W^{\gamma,\infty}([-1,1]^d)$, there exists a $g_\epsilon \in \mathcal{M}^{\mathrm{DNN}}$ and $c_1 > 0$, such that $\|g_\epsilon - g^*\|_\infty \leq \epsilon$.*

2. *$H(\epsilon) \leq c_1(\log(\frac{1}{\epsilon}) + 1)$ and $W(\epsilon), U(\epsilon) \leq c_1 \cdot \epsilon^{-\frac{d}{\gamma}}(\log(\frac{1}{\epsilon}) + 1)$, where H denotes the number of layers, W the number of weights, and U the total number of hidden units.*

The next result is Lemma 3 in Zhou et al. (2023), which establishes the approximation error bound for the MDN. Note that our Assumptions 7 and 8 are analogous to Assumptions 2 and 3 from Zhou et al. (2023). In the following result, f^* is the true density.

Lemma 7. *Suppose Assumptions 2 and 3 from Zhou et al. (2023) hold. Then for any $0 < \epsilon \leq 1$ and integer $G \geq 1$, there exists a set of DNN functions $\{(g_{g1}, g_{g2}, g_{g3})_{g=1}^G\}$ whose network architectures depend on ϵ such that*

1. *$\|f^* - f_T\|_\infty \leq C_2' G^{4\omega_2}\epsilon + O(G^{-\omega_1})$ where $f_T = \sum_{g=1}^G g_{g1}(x) \frac{1}{\sqrt{2\pi} g_{g3}(x)} e^{-\frac{(y - g_{g2}(x))^2}{2g_{g3}^2(x)}}$, for some constant $C_2' > 0$ that is independent of ϵ;*

2. *the number of hidden layers H_{gj}, the number of parameters W_{gj}, and the total number of hidden units U_{gj} satisfy*

   (a) $H_{gj} \leq c_1\{\log(G/\epsilon) + 1\}$

   (b) $W_{gj} \leq c_1 G^{\frac{d}{\gamma}}\ \epsilon^{-\frac{d}{\gamma}}\{\log(G/\epsilon) + 1\}$

   (c) $U_{gj} \leq c_1 G^{\frac{d}{\gamma}}\ \epsilon^{-\frac{d}{\gamma}}\{\log(G/\epsilon) + 1\}$

The next result is Lemma 4 from Zhou et al. (2023), which connects the metric entropy between the MDN function class and DNN function class.

Lemma 8. *Suppose Assumptions 2 and 3 from Zhou et al. (2023) hold. Then there exists a constant $c_3 > 0$, such that*

$$\log \mathcal{T}(\delta, \mathcal{M}, \|\cdot\|_n) \leq G \log \mathcal{T}\left(\frac{\delta}{c_3 G^{3/2+3\omega_2}}, \mathcal{M}^{\mathrm{DNN}}, \|\cdot\|_n\right),$$

where G is the number of mixture components.

The following result is based on Lemma 6 in Farrell et al. (2021) (Theorem 6 in Bartlett et al. (2019) in the ReLU case) which is used to bound the complexity of the function class of DNNs. Note that Pdim denotes the pseudo VC dimension of the function class.

Lemma 9. *Consider a ReLU network architecture $\mathcal{M} = \mathcal{M}^{\mathrm{DNN}}(W, L, U)$, then the pseudo VC dimension satisfies*

$$\mathrm{Pdim}(\mathcal{M}) \leq C \cdot WL \log W,$$

for some universal constant $C > 0$.

The next result is Lemma 5 in Zhou et al. (2023).

Lemma 10. *Suppose there is a sequence of stationary time series, $U = (X_1, X_2, \ldots, X_T)$, with a common marginal distribution, and each $X_t \in \mathbb{R}^d$, $t = 1, \ldots, T$. Suppose the β-mixing coefficient of the sequence satisfies that $\beta(p) \leq c_1 e^{-c_2 p}$ for some c_1, c_2. Then there exists a sequence, $U_i^0 = (X_{ip+1}^0, X_{ip+2}^0, \ldots, X_{ip+p}^0)$, for $i \geq 0$, such that*

1. *U_i^0 has the same distribution as $U_i = (X_{ip+1}, X_{ip+2}, \ldots, X_{ip+p})$.*
2. *The sequence $\{U_{2i}^0\}_{i \geq 0}$ is i.i.d., and so is $\{U_{2i+1}^0\}_{i \geq 0}$.*
3. *For any $i \geq 0$, $\mathbb{P}(U_i \neq U_i^0) \leq \beta(p)$*
4. *$\mathbb{P}(|G_T(U) - G_T(U^0)| \neq 0) \leq p^{-1} T \beta(p)$, where $G_T(U) = \sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^T f(X_{t+1}|X_t) - \mathbb{E}f\right\}$, and $G_T(U^0) = \sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^T f(X_{t+1}^0|X_t^0) - \mathbb{E}f\right\}$.*

C Additional Results

C.1 Mixture Density Network

The following result extends the guarantees for the mixture density network (MDN) estimator from Zhou et al. (2023) to the setting of covariates with lower-dimensional support. We must consider this extension because the predictive state embeddings Z_t take values in $\Phi(\mathcal{S}) \subset \mathbb{R}^k$ but the predictive state space $\mathcal{S} \subset \mathbb{R}^r$ has Hausdorff dimension p, where $k \geq 2p + 1$. In particular, we do not assume $P_{\tilde{Z}_{t,j}}$ is absolutely continuous with respect to the Lebesgue measure. Originally, Zhou et al. (2023) only needed to consider estimating the conditional distributions of X_{t+1} given X_t and X_{t-1} given X_t, and they did not need to consider covariates with lower-dimensional support.

Recall the notation from Section 3.2. Also, for the following result, denote the supports of $\tilde{Z}_{t,j}$ and $X_{t+1,j}$ by $\tilde{\mathcal{Z}}_j$ and $\mathcal{X}_j$, respectively, the Lebesgue measure restricted to $\mathcal{X}_j$ by $\lambda_{\mathcal{X}_j}$, and the distribution of $\tilde{Z}_{t,j}$ by $P_{\tilde{Z}_{t,j}}$. Note that we use the superscript "oracle" to distinguish the estimators, $\hat{f}_j^{\mathrm{oracle}}$, $j \in [d]$, in Theorem 6 which use the true covariates $\tilde{Z}_{t,j}$, $j \in [d]$, from the feasible estimators $\hat{f}_j$, $j \in [d]$, in Section 3.2 which use the estimated covariates $\hat{\tilde{Z}}_{t,j}$, $j \in [d]$. When the covariates do not need to be estimated, the "oracle" in the superscript can be ignored. For the following result, denote the smallest and largest number of mixture components, respectively, by

$$M_- = \min_{j \in [d]}(M_j), \quad M_+ = \max_{j \in [d]}(M_j).$$

Theorem 6. *Suppose Assumptions 3, 7, 9, 11 hold. Then, there exist MDN function classes $\mathcal{M}_{n,j}^{\mathrm{MDN}}$, $j \in [d]$, satisfying Assumption 8, such that the oracle conditional joint density estimator*

$$\hat{f}^{\mathrm{oracle}} = \prod_{j=1}^d \hat{f}_j^{\mathrm{oracle}}, \quad \hat{f}_j^{\mathrm{oracle}} \in \operatorname*{arg\,max}_{f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}}} \sum_{t=m+1}^{n-m-1} \log f_j\left(X_{t+1,j} \mid \tilde{Z}_{t,j}\right), \quad j = 1, \ldots, d, \quad (35)$$

satisfies

$$\begin{aligned}&\left\|\hat{f}^{\text{oracle}} - f_{X_{t+1}|Z_t}\right\|_{L^2} \\ &= \left(\int_{\mathcal{Z}}\int_{\mathcal{X}} \left|\hat{f}^{\text{oracle}}(x \mid z) - f_{X_{t+1}|Z_t}(x \mid z)\right|^2 dx dP_{Z_t}(z)\right)^{1/2} \\ &\leq c\left(M_-^{-\omega_1} + M_+^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM_+)\right),\end{aligned} \tag{36}$$

for some constant $c > 0$, with probability at least $1 - O(n^{-1})$.

C.2 Predictive States

The next result generalizes Corollary 2 of Theorem 1 in Shalizi & Crutchfield (2001) from discrete measurable spaces to standard Borel spaces. First, we introduce the general setting.

Let $(\mathcal{X}, \mathfrak{X})$ and $(\mathcal{S}, \mathfrak{S})$ be standard Borel spaces. Consider the model

$$X_{t+1} = G(\varepsilon_{t+1}, S_t), \quad S_t = \tau(S_{t-1}, X_t) = R(X_{-\infty:t}), \quad t \in \mathbb{Z}, \tag{37}$$

where X_t is an $(\mathcal{X}, \mathfrak{X})$-valued sequence, ε_t, $t \in \mathbb{Z}$, is an iid sequence of random elements satisfying $\varepsilon_{t+1} \perp\!\!\!\perp \mathcal{F}_t$ for all $t \in \mathbb{Z}$, S_t, $t \in \mathbb{Z}$, is an $(\mathcal{S}, \mathfrak{S})$-valued sequence, and the functions G, τ, and R are each Borel measurable. The conditional distribution of X_{t+1} given S_t is given by

$$\mathcal{L}(X_{t+1} \mid S_t) = \mathcal{L}(X_{t+1} \mid \mathcal{F}_t), \quad t \in \mathbb{Z}, \tag{38}$$

because $S_t = R(X_{-\infty:t})$ is $\mathcal{F}_t$-measurable and because $\varepsilon_{t+1} \perp\!\!\!\perp \mathcal{F}_t$. By Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)), equation 38 is equivalent to one-step-ahead predictive sufficiency,

$$X_{t+1} \perp\!\!\!\perp X_{-\infty:t} \mid S_t, \quad t \in \mathbb{Z}. \tag{39}$$

Theorem 7. *The predictive states from equation 37 have the strong predictive sufficiency property,*

$$X_{t+1:\infty} \perp\!\!\!\perp X_{-\infty:t} \mid S_t, \quad t \in \mathbb{Z}. \tag{40}$$

D Proofs of Additional Results

D.1 Proof of Theorem 6

The proof is broken into three main steps, as in the proof of Theorem 3 in Zhou et al. (2023).

To begin, we introduce some notation. Recall that

$$\tilde{Z}_{t,j} = (Z_t^\top, X_{t+1,1}, \ldots, X_{t+1,j-1})^\top, \quad j \in [d].$$

Denote the sample size by

$$N = n - 2m - 1,$$

where n is the number of observations of X_t, $t = 1, \ldots, n$, and m is from the random Fourier features from equation 10.

For every measurable function $h = h(X_{t+1,j}, \tilde{Z}_{t,j})$, define

$$P_j h = \mathbb{E}[h(X_{t+1,j}, \tilde{Z}_{t,j})], \quad P_{n,j} h = \frac{1}{N}\sum_{t=m+1}^{n-m-1} h(X_{t+1,j}, \tilde{Z}_{t,j}), \tag{41}$$

and let

$$\|h\|_{L^2,j}^2 = \int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} |h(x_j, \tilde{z}_j)|^2 \, dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j), \quad \|h\|_{n,j}^2 = \frac{1}{N}\sum_{t=m+1}^{n-m-1} \left|h(X_{t+1,j}, \tilde{Z}_{t,j})\right|^2.$$

Similarly, denote

$$f_j = f_j(X_{t+1,j} \mid \tilde{Z}_{t,j}),$$

for some conditional density $f_j(\cdot \mid \cdot)$, so that we may concisely write

$$P_j f_j = \mathbb{E}[f_j(X_{t+1,j} \mid \tilde{Z}_{t,j})],$$

and similarly $P_j \log f_j = \mathbb{E}[\log f_j(X_{t+1,j} \mid \tilde{Z}_{t,j})]$.

Step 1. This step corresponds to Step 1 in the proof of Theorem 3 in Zhou et al. (2023), which itself uses ideas from Farrell et al. (2021). In this step, the observations are temporarily regarded as independent, and the extension to dependent observations is in Step 2.

Step 1.1. This step corresponds to Step 1.1 in the proof of Theorem 3 in Zhou et al. (2023), which starts in the iid setting.

We derive upper and lower bounds for the quantity $K_j(\hat{f}_j^{\text{oracle}})$, where for any measurable function h that is P_j-almost surely positive and for which the following expectation is finite, we define

$$\begin{aligned} K_j(h) &= P_j\left(\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log h\right) \\ &= \int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) - \log h(x_j \mid \tilde{z}_j)\right] dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j). \end{aligned} \tag{42}$$

For $a \in [0,1]$, define

$$f_{j,a}(x_j \mid \tilde{z}_j) = (1-a) f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) + a f_j(x_j \mid \tilde{z}_j).$$

Note that $f_{j,1}(x_j \mid \tilde{z}_j) = f_j(x_j \mid \tilde{z}_j)$ and $f_{j,0}(x_j \mid \tilde{z}_j) = f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)$, and that $f_{j,a}$ need not belong to the function class $\mathcal{M}_{n,j}^{\text{MDN}}$.

By Taylor's theorem (with the integral form of the remainder) applied to $a \mapsto K_j(f_{j,a})$ about $a = 0$ then evaluating at $a = 1$, we have

$$K_j(f_j) = K_j(f_{j,1}) = K_j(f_{j,0}) + \frac{d}{da}K_j(f_{j,a})\bigg|_{a=0} + \int_0^1 (1-a)\frac{d^2}{da^2}K_j(f_{j,a})da, \tag{43}$$

where $K_j(f_{j,0}) = 0$ by equation 42. Explicitly, we have

$$\begin{aligned} &K_j(f_j) \\ &= -\int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} \frac{f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)}{f_{j,0}(x_j \mid \tilde{z}_j)}\left[f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\right] dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \\ &+ \int_0^1 (1-a)\int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} \frac{f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)}{f_{j,a}^2(x_j \mid \tilde{z}_j)}\left|f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\right|^2 dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j)da \\ &= -\int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j}\left[f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\right] dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) + \\ &\int_0^1 (1-a)\int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} \frac{f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)}{f_{j,a}^2(x_j \mid \tilde{z}_j)}\left|f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\right|^2 dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j)da \\ &= \int_{\tilde{\mathcal{Z}}_j}\int_{\mathbb{R}\setminus\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) + \\ &\int_0^1 (1-a)\int_{\tilde{\mathcal{Z}}_j}\int_{\mathcal{X}_j} \frac{f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)}{f_{j,a}^2(x_j \mid \tilde{z}_j)}\left|f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)\right|^2 dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j)da, \end{aligned}$$

where the first equality is by differentiating equation 42 along the path $f_{j,a}$ and applying equation 43, the second equality is by cancellation because $f_{j,0}(x_j \mid \tilde{z}_j) = f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j)$ as noted above,

and the third equality is because

$$
\begin{aligned}
&- \int_{\tilde{\mathcal{Z}}_j} \int_{\mathcal{X}_j} \left[f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \right] dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \\
&= \int_{\tilde{\mathcal{Z}}_j} \left[\int_{\mathcal{X}_j} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) dx_j - \int_{\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j \right] dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \\
&= \int_{\tilde{\mathcal{Z}}_j} \left[1 - \int_{\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j \right] dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \\
&= \int_{\tilde{\mathcal{Z}}_j} \left[\left(\int_{\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j + \int_{\mathbb{R}\setminus\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j \right) - \int_{\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j \right] dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \\
&= \int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j),
\end{aligned}
$$

where the first equality follows by linearity, the second because $f_{X_{t+1,j}|\tilde{Z}_{t,j}}(\cdot \mid \tilde{z}_j)$ integrates to one over the support $\mathcal{X}_j$, and the third because $f_j(\cdot \mid \tilde{z}_j)$ integrates to one over $\mathbb{R}$, and the final equality follows by canceling terms.

Denote $c_- = \min(c_f^-, c_{\text{MDN}}^{-1})$ and $c_+ = \max(c_f^+, c_{\text{MDN}})$, where the constants $c_f^-, c_f^+, c_{\text{MDN}}$ are from Assumptions 7 and 8. Note that $f_{j,a} \in [c_-, c_+]$, so we may upper and lower bound $K_j(f_j)$ by

$$
K_j(f_j) \leq \int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) + \frac{c_f^+}{2c_-^2} \left\| f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2, \tag{44}
$$

$$
K_j(f_j) \geq \int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_j(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) + \frac{c_f^-}{2c_+^2} \left\| f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2. \tag{45}
$$

Therefore, for $\hat{f}_j^{\text{oracle}}$ from the theorem statement and for any DNN approximation in the MDN function class, $f_j^{\text{DNN}} \in \mathcal{M}_{n,j}^{\text{MDN}}$, we can upper and lower bound $K_j(\hat{f}_j^{\text{oracle}})$ by

$$
\begin{aligned}
&K_j(\hat{f}_j^{\text{oracle}}) \leq (P_j - P_{n,j}) \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}_j^{\text{oracle}} \right] + P_{n,j} \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_j^{\text{DNN}} \right], \\
&K_j(\hat{f}_j^{\text{oracle}}) \geq \frac{c_f^-}{2c_+^2} \left\| \hat{f}_j^{\text{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2,
\end{aligned}
$$

where the lower bound follows by dropping the nonnegative first term from equation 45, and the upper bound follows by adding and subtracting the same term $P_{n,j}[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}_j^{\text{oracle}}]$ to equation 42, where $P_{n,j}$ is from equation 41, linearity, the triangle inequality, and noting that $\hat{f}_j^{\text{oracle}}$ from equation 35 is the argmax so we may upper bound by any $f_j^{\text{DNN}} \in \mathcal{M}_{n,j}^{\text{MDN}}$.

Step 1.2. This step corresponds to Step 1.2 in the proof of Theorem 3 in Zhou et al. (2023). We apply the approximation result from Lemma 3 in Zhou et al. (2023).

Fix the dimension $j \in [d]$, number of mixture components $M_j \in \mathbb{N}$, and the approximation level $\epsilon \in (0, 1/4]$. By Assumption 7, the conditional density $f_{X_{t+1,j}|\tilde{Z}_{t,j}}$ is approximated by the mixture

$$
\sup_{x_j, \tilde{z}_j} \left| f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) - \sum_{g=1}^{M_j} \frac{\alpha_{g,j}(\tilde{z}_j)}{\sqrt{2\pi}\sigma_{g,j}(\tilde{z}_j)} \exp\left(-\frac{(x_j - \mu_{g,j}(\tilde{z}_j))^2}{2\sigma_{g,j}^2(\tilde{z}_j)} \right) \right| \leq c_{\text{gmix}} M_j^{-\omega_1}. \tag{46}
$$

Step 1.2.1. Next, we verify that the DNN approximation result used in Lemma 3 of Zhou et al. (2023) applies to the functions in equation 46. By Assumption 7, for each $g \in [M_j]$, the functions $\alpha_{g,j}(\cdot)$, $\mu_{g,j}(\cdot)$, and $\sigma_{g,j}(\cdot)$ admit γ-Sobolev-smooth extensions to the domain

$$
B_j = [-b_{\mathcal{X}}, b_{\mathcal{X}}]^{k+j-1}.
$$

We may map B_j onto $[-1,1]^{k+j-1}$, i.e., by normalizing $B_j/b_{\mathcal{X}} = [-1,1]^{k+j-1}$, so that the Sobolev norms of the functions will remain bounded by a constant that depends on $b_{\mathcal{X}}$, γ, and c_{Sob}. Hence, Lemma 6 (Lemma 2 in Zhou et al. (2023)) applies with input dimension given by $k+j-1$.

Let $\ell = 1,2,3$ correspond to $\alpha_{g,j}$, $\mu_{g,j}$, $\sigma_{g,j}$, respectively. For each $g \in [M_j]$ and $\ell \in \{1,2,3\}$, let

$$H_{g,j,\ell}(\epsilon), \quad W_{g,j,\ell}(\epsilon), \quad U_{g,j,\ell}(\epsilon),$$

denote the number of hidden layers, the number of weights, and the total number of hidden units in the deep neural network (DNN) used to approximate the gth mixture coefficient function.

Lemma 6 (Lemma 2 in Zhou et al. (2023)) with approximation level ϵ/M_j yields the approximation

$$\left\|\alpha_{g,j}^{\text{DNN}} - \alpha_{g,j}\right\|_\infty \leq \frac{\epsilon}{M_j}, \quad \left\|\mu_{g,j}^{\text{DNN}} - \mu_{g,j}\right\|_\infty \leq \frac{\epsilon}{M_j}, \quad \left\|\sigma_{g,j}^{\text{DNN}} - \sigma_{g,j}\right\|_\infty \leq \frac{\epsilon}{M_j}. \tag{47}$$

Note that the true mean and scale satisfy the bounds from Assumption 7, so we may clip the DNN approximations to the ranges from Assumption 8 without increasing the approximation errors in equation 47.

For a $(k+j-1)$-dimensional γ-Sobolev-smooth function approximated with accuracy δ, Lemma 6 (Lemma 2 in Zhou et al. (2023)) yields the following bounds on the number of hidden layers, weights, and hidden units,

$$H(\delta) \leq C(\log(1/\delta)+1), \quad W(\delta), U(\delta) \leq C\delta^{-(k+j-1)/\gamma}(\log(1/\delta)+1).$$

Letting $\delta = \epsilon/M_j$ yields

$$\delta^{-(k+j-1)/\gamma} = M_j^{(k+j-1)/\gamma}\epsilon^{-(k+j-1)/\gamma},$$

so there exists a constant $C > 0$, independent of j, M_j, and ϵ, such that, for all $\ell \in \{1,2,3\}$,

$$\begin{aligned}
H_{g,j,\ell}(\epsilon) &\leq C(\log\left(\frac{M_j}{\epsilon}\right)+1), \\
W_{g,j,\ell}(\epsilon) &\leq CM_j^{(k+j-1)/\gamma}\epsilon^{-(k+j-1)/\gamma}(\log\left(\frac{M_j}{\epsilon}\right)+1), \\
U_{g,j,\ell}(\epsilon) &\leq CM_j^{(k+j-1)/\gamma}\epsilon^{-(k+j-1)/\gamma}(\log\left(\frac{M_j}{\epsilon}\right)+1).
\end{aligned} \tag{48}$$

Note that Lemma 7 (Lemma 3 in Zhou et al. (2023)) applies because Assumption 7 gives the same uniform bounds on the functions in the Gaussian mixture, and Assumption 3 implies $\sup_{x_j \in \mathcal{X}_j} |x_j| < \infty$. Also, note that an inspection of the proof of Lemma 7 shows that replacing the bound $|x_j| \leq 1$ by $|x_j| \leq b_{\mathcal{X}}$ only changes the constants.

Next, we introduce the approximating MDN, $f_{j,\epsilon}^{\text{app}}$, which is obtained by replacing the (infeasible) functions $\alpha_{g,j}(\cdot)$, $\mu_{g,j}(\cdot)$, and $\sigma_{g,j}(\cdot)$ from Assumption 7 by the DNN approximations. Specifically,

$$f_{j,\epsilon}^{\text{app}}(x_j \mid \tilde{z}_j) = \sum_{g=1}^{M_j} \alpha_{g,j}^{\text{app}}(\tilde{z}_j) \frac{1}{\sqrt{2\pi}\sigma_{g,j}^{\text{DNN}}(\tilde{z}_j)} \exp\left(-\frac{\left(x_j - \mu_{g,j}^{\text{DNN}}(\tilde{z}_j)\right)^2}{2\left(\sigma_{g,j}^{\text{DNN}}(\tilde{z}_j)\right)^2}\right), \tag{49}$$

where $\alpha_{g,j}^{\text{app}}(\cdot)$ are the normalized DNN approximations of the mixture weights defined as

$$\alpha_{g,j}^{\text{app}} = \frac{a_{g,j}^{\text{clip}}}{\sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}}, \quad a_{g,j}^{\text{clip}} = \max(\alpha_{g,j}^{\text{DNN}} - \frac{\epsilon}{M_j}, 0). \tag{50}$$

By equation 47, we have

$$0 \leq a_{g,j}^{\text{clip}} \leq \alpha_{g,j}, \quad \alpha_{g,j} - a_{g,j}^{\text{clip}} \leq \frac{2\epsilon}{M_j}. \tag{51}$$

Since the mixture weights sum to one, we have, by equation 51, summing over the mixture components, and subtraction, that

$$1 - 2\epsilon \leq \sum_{g=1}^{M_j} a_{g,j}^{\text{clip}} \leq 1. \tag{52}$$

Next, we will bound the quantity

$$\sum_{g=1}^{M_j} \left|\alpha_{g,j}^{\text{app}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right|. \tag{53}$$

By the triangle inequality, we have

$$\sum_{g=1}^{M_j} \left|\alpha_{g,j}^{\text{app}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right| \leq \sum_{g=1}^{M_j} \left|\alpha_{g,j}^{\text{app}}(\tilde{z}_j) - a_{g,j}^{\text{clip}}(\tilde{z}_j)\right| + \sum_{g=1}^{M_j} \left|a_{g,j}^{\text{clip}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right|, \tag{54}$$

so we will consider each term separately.

Step 1.2.1.1. For the first term in equation 54, since

$$\alpha_{g,j}^{\text{app}}(\tilde{z}_j) = \frac{a_{g,j}^{\text{clip}}(\tilde{z}_j)}{\sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}(\tilde{z}_j)},$$

and

$$\sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}(\tilde{z}_j) \leq 1,$$

we have

$$\sum_{g=1}^{M_j} \left|\alpha_{g,j}^{\text{app}}(\tilde{z}_j) - a_{g,j}^{\text{clip}}(\tilde{z}_j)\right| \leq 2\epsilon, \tag{55}$$

because, by distributing terms and rearranging the terms from equation 52, we have

$$\sum_{g=1}^{M_j} \left|\alpha_{g,j}^{\text{app}}(\tilde{z}_j) - a_{g,j}^{\text{clip}}(\tilde{z}_j)\right| = \sum_{g=1}^{M_j} a_{g,j}^{\text{clip}}(\tilde{z}_j) \left(\frac{1}{\sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}(\tilde{z}_j)} - 1\right) = 1 - \sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}(\tilde{z}_j) \leq 2\epsilon.$$

Step 1.2.1.2. For the second term in equation 54, we must consider two cases.

First, if $\alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) \geq \epsilon/M_j$, then

$$\left|a_{g,j}^{\text{clip}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right| = \frac{\epsilon}{M_j},$$

since $a_{g,j}^{\text{clip}}(\tilde{z}_j) = \max\left(\alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) - \frac{\epsilon}{M_j}, 0\right)$.

Second, if $\alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) < \epsilon/M_j$, then

$$a_{g,j}^{\text{clip}}(\tilde{z}_j) = 0. \tag{56}$$

Since $\alpha_{g,j}(\tilde{z}_j) \geq 0$ and, by equation 47, we have

$$\left|\alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) - \alpha_{g,j}(\tilde{z}_j)\right| \leq \frac{\epsilon}{M_j},$$

then we must have

$$\alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) \geq -\frac{\epsilon}{M_j}. \tag{57}$$

Therefore, by equation 56 and equation 57, we have

$$\left|a_{g,j}^{\text{clip}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right| \leq \left|0 + \frac{\epsilon}{M_j}\right| = \frac{\epsilon}{M_j}.$$

Hence, in both cases, we have

$$\left|a_{g,j}^{\text{clip}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j)\right| \leq \frac{\epsilon}{M_j},$$

so summing over the mixture components yields

$$\sum_{g=1}^{M_j} \left| a_{g,j}^{\text{clip}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) \right| \leq \epsilon. \tag{58}$$

Combining equation 54, equation 55, and equation 58, we obtain

$$\sum_{g=1}^{M_j} \left| \alpha_{g,j}^{\text{app}}(\tilde{z}_j) - \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) \right| \leq 3\epsilon. \tag{59}$$

Step 1.2.1.3. By equation 47, we have the DNN approximation

$$\sum_{g=1}^{M_j} \left| \alpha_{g,j}^{\text{DNN}}(\tilde{z}_j) - \alpha_{g,j}(\tilde{z}_j) \right| \leq M_j \frac{\epsilon}{M_j} = \epsilon. \tag{60}$$

Therefore, by the triangle inequality, equation 59, and equation 60, we have

$$\sum_{g=1}^{M_j} \left| \alpha_{g,j}^{\text{app}}(\tilde{z}_j) - \alpha_{g,j}(\tilde{z}_j) \right| \leq 3\epsilon + \epsilon = 4\epsilon.$$

Note that the quantity below is bounded,

$$\sup_{x_j, \tilde{z}_j, g} \frac{1}{\sqrt{2\pi}\sigma_{g,j}^{\text{DNN}}(\tilde{z}_j)} \exp\left(-\frac{(x_j - \mu_{g,j}^{\text{DNN}}(\tilde{z}_j))^2}{2(\sigma_{g,j}^{\text{DNN}}(\tilde{z}_j))^2} \right) \leq C M_j^{\omega_2}.$$

By equation 59 and the previous bound on the components of the Gaussian mixture, replacing the DNN mixture weights $\alpha_{g,j}^{\text{DNN}}$ by the normalized weights $\alpha_{g,j}^{\text{app}}$ introduces an additional error of at most

$$C M_j^{\omega_2} \epsilon, \tag{61}$$

for some constant $C > 0$.

Therefore, by the triangle inequality, Lemma 7 (with input dimension $k + j - 1$), and the approximation bound from equation 61, there exists an MDN $f_{j,\epsilon}^{\text{app}}$ from equation 49 satisfying

$$\sup_{x_j \in \mathcal{X}_j, \tilde{z}_j \in \tilde{\mathcal{Z}}_j} \left| f_{j,\epsilon}^{\text{app}}(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \right| \leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}). \tag{62}$$

Note that, for sufficiently small ϵ and sufficiently large c_{MDN}, the construction of $f_{j,\epsilon}^{\text{app}}$ from equation 49 belongs to the function class $\mathcal{M}_{n,j}^{\text{MDN}}$ from Assumption 8.

Step 1.2.2. Next, we check that the architecture above yields the parameter count in Assumption 8. There are $3M_j$ networks in the jth MDN, i.e., one network for each of $\alpha_{g,j}$, $\mu_{g,j}$, and $\sigma_{g,j}$, $g \in [M_j]$. Therefore, denoting the total number of parameters in the jth MDN by $N_{j,\epsilon}^{\text{par}}$, we have

$$N_{j,\epsilon}^{\text{par}} \leq C M_j \max_{\substack{g \in [M_j] \\ \ell \in \{1,2,3\}}} (W_{g,j,\ell}(\epsilon) + U_{g,j,\ell}(\epsilon)) \tag{63}$$

$$\leq C M_j^{(k+j-1+\gamma)/\gamma} \epsilon^{-(k+j-1)/\gamma} (\log\left(\frac{M_j}{\epsilon}\right) + 1). \tag{64}$$

Letting

$$\epsilon = n^{-\gamma/(2\gamma+k+j-1)},$$

so that

$$\epsilon^{-(k+j-1)/\gamma} = n^{(k+j-1)/(2\gamma+k+j-1)}, \quad \log(M_j/\epsilon) \leq C \log(nM_j),$$

we have

$$N_{j,\epsilon}^{\text{par}} \leq C M_j^{(k+j-1+\gamma)/\gamma} n^{(k+j-1)/(2\gamma+k+j-1)} \log(nM_j), \tag{65}$$

as required by Assumption 8.

Step 1.2.3. Next, we derive two upper bounds that will be used later.

First, we upper bound the quantity

$$\int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_{j,\epsilon}^{\text{app}}(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j),$$

where the approximating MDN $f_{j,\epsilon}^{\text{app}}$ is from equation 49.

For each $g \in [M_j]$ and $\tilde{z}_j \in \tilde{\mathcal{Z}}_j$, let

$$P_{g,j,\tilde{z}_j} = N\left(\mu_{g,j}(\tilde{z}_j), \sigma_{g,j}^2(\tilde{z}_j)\right),$$

denote the Gaussian distribution corresponding to the mixture from Assumption 7, and let

$$P_{g,j,\tilde{z}_j}^{\text{app}} = N\left(\mu_{g,j}^{\text{DNN}}(\tilde{z}_j), (\sigma_{g,j}^{\text{DNN}})^2(\tilde{z}_j)\right),$$

denote the corresponding Gaussian distribution obtained by replacing $\mu_{g,j}$ and $\sigma_{g,j}$ with their DNN approximations from equation 47.

Denote the squared Hellinger distance between two distributions P and Q, with densities p and q, by

$$H^2(P,Q) = \int_{\mathbb{R}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

By the reverse triangle inequality, we have

$$\sqrt{P_{g,j,\tilde{z}_j}^{\text{app}}(\mathbb{R}\setminus\mathcal{X}_j)} \le \sqrt{P_{g,j,\tilde{z}_j}(\mathbb{R}\setminus\mathcal{X}_j)} + H(P_{g,j,\tilde{z}_j}^{\text{app}}, P_{g,j,\tilde{z}_j}).$$

Squaring the previous inequality and using the inequality $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$P_{g,j,\tilde{z}_j}^{\text{app}}(\mathbb{R}\setminus\mathcal{X}_j) \le 2P_{g,j,\tilde{z}_j}(\mathbb{R}\setminus\mathcal{X}_j) + 2H^2\left(P_{g,j,\tilde{z}_j}^{\text{app}}, P_{g,j,\tilde{z}_j}\right).$$

The squared Hellinger distance between Gaussians (which is known in closed form), and the lower bounds

$$\sigma_{g,j}(\tilde{z}_j) \ge c_\sigma^- M_j^{-\omega_2}, \quad \sigma_{g,j}^{\text{DNN}}(\tilde{z}_j) \ge c_{\text{MDN}}^{-1} M_j^{-\omega_2},$$

from Assumptions 7 and 8, and equation 47 yield

$$H^2(P_{g,j,\tilde{z}_j}^{\text{app}}, P_{g,j,\tilde{z}_j}) \le C M_j^{2\omega_2}\left(\frac{\epsilon}{M_j}\right)^2.$$

By equation 50 and equation 52, we have

$$\alpha_{g,j}^{\text{app}} = \frac{a_{g,j}^{\text{clip}}}{\sum_{h=1}^{M_j} a_{h,j}^{\text{clip}}} \le \frac{a_{g,j}^{\text{clip}}}{1-2\epsilon} \le \frac{\alpha_{g,j}}{1-2\epsilon}.$$

Using the corresponding DNN approximation of the mixture weights, which satisfy

$$\alpha_{g,j}^{\text{app}} \le 2\alpha_{g,j}, \tag{66}$$

because $1/(1-2\epsilon) \in (1,2]$ since $\epsilon \in (0, 1/4]$, summing over $g \in [M_j]$, integrating over $\tilde{z}_j$, and applying the upper bounds from Assumption 9 yields

$$\int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_{j,\epsilon}^{\text{app}}(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \le C M_j^{-2\omega_1} + C M_j^{2\omega_2 - 2}\epsilon^2,$$

where $f_{j,\epsilon}^{\text{app}}$ is from equation 49. Since $M_j \ge 1$ and $\omega_2 > 0$, we have

$$M_j^{2\omega_2-2}\epsilon^2 \le M_j^{8\omega_2}\epsilon^2.$$

Therefore, we have the desired upper bound

$$\int_{\tilde{\mathcal{Z}}_j} \int_{\mathbb{R}\setminus\mathcal{X}_j} f_{j,\epsilon}^{\text{app}}(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j) \le C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})^2. \tag{67}$$

Second, we upper bound the quantity from equation 44 in Step 1.1. By equation 62 and the compactness of $\mathcal{X}_j$, we have

$$\left\| f_{j,\epsilon}^{\text{app}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2 \leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})^2, \tag{68}$$

where $f_{j,\epsilon}^{\text{app}}$ is from equation 49. By equation 67 and equation 68, the quantity from equation 44 in Step 1.1 can be bounded by

$$K_j(f_{j,\epsilon}^{\text{app}}) \leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})^2. \tag{69}$$

Step 1.2.4. Next, we obtain the final approximation bound as in Step 1.2 of Zhou et al. (2023). By Assumptions 7 and 8, both densities are uniformly bounded away from zero on $\mathcal{X}_j \times \tilde{\mathcal{Z}}_j$, so the mean value theorem applied to the logarithm and equation 62 imply that

$$|\log f_{X_{t+1,j}|\tilde{Z}_{t,j}}(X_{t+1,j} \mid \tilde{Z}_{t,j}) - \log f_{j,\epsilon}^{\text{app}}(X_{t+1,j} \mid \tilde{Z}_{t,j})| \leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}),$$

almost surely, and that

$$\text{Var}(\log f_{X_{t+1,j}|\tilde{Z}_{t,j}}(X_{t+1,j} \mid \tilde{Z}_{t,j}) - \log f_{j,\epsilon}^{\text{app}}(X_{t+1,j} \mid \tilde{Z}_{t,j})) \leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})^2.$$

Recall $P_{n,j}$ from equation 41. Under the setting of independent observations from Step 1, Bernstein's inequality gives, for every $\tilde{r} > 0$, with probability at least $1 - \exp(-\tilde{r})$,

$$\begin{aligned} &P_{n,j}\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_{j,\epsilon}^{\text{app}}\right] \\ &\leq C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})^2 + C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})\sqrt{\frac{\tilde{r}}{N}} + C(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1})\frac{\tilde{r}}{N}. \end{aligned} \tag{70}$$

Step 1.3. This step corresponds to Step 1.3 in the proof of Theorem 3 in Zhou et al. (2023).

We bound the quantity

$$(P_j - P_{n,j})\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}_j^{\text{oracle}}\right],$$

from Step 1.1 using arguments analogous to those from Zhou et al. (2023) who themselves use the ideas from Farrell et al. (2021).

Step 1.3.1. For a constant $r_{0,j} > 0$, define

$$\mathcal{M}_{n,j}^{\text{MDN}}(r_{0,j}) = \left\{ f_j \in \mathcal{M}_{n,j}^{\text{MDN}} : \left\| f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j} \leq r_{0,j} M_j^{\omega_2} \right\}.$$

Note that, since $\mathcal{X}_j$ is compact and both densities are uniformly bounded by c_+, we can select $r_{0,j} > 0$ sufficiently large so that

$$\left\| f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j} \leq r_{0,j} M_j^{\omega_2},$$

for every $f_j \in \mathcal{M}_{n,j}^{\text{MDN}}$, since $M_j^{\omega_2} \geq 1$. Define the corresponding classes of differences by

$$\begin{aligned} \mathcal{H}_j(r_{0,j}) &= \left\{ f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} : f_j \in \mathcal{M}_{n,j}^{\text{MDN}}(r_{0,j}) \right\} \cup \{0\}, \\ \mathcal{G}_j(r_{0,j}) &= \left\{ \log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_j : f_j \in \mathcal{M}_{n,j}^{\text{MDN}}(r_{0,j}) \right\} \cup \{0\}. \end{aligned}$$

Recall c_- and c_+ from Step 1.1. By Assumption 7, we have

$$c_f^- \|h\|_{L^2,j}^2 \leq P_j h^2 \leq c_f^+ \|h\|_{L^2,j}^2, \tag{71}$$

for any measurable function h, because

$$P_j h^2 = \int_{\tilde{\mathcal{Z}}_j} \int_{\mathcal{X}_j} |h(x_j, \tilde{z}_j)|^2 f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) dx_j dP_{\tilde{Z}_{t,j}}(\tilde{z}_j),$$

Thus, by the mean value theorem for the logarithm, we have

$$\left| \log f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) - \log f_j(x_j \mid \tilde{z}_j) \right| \leq c_-^{-1} \left| f_j(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \right|,$$

so

$$\sup_{g_j \in \mathcal{G}_j(r_{0,j})} P_j g_j^2 \leq \frac{c_f^+}{c_-^2} r_{0,j}^2 M_j^{2\omega_2}, \qquad \sup_{g_j \in \mathcal{G}_j(r_{0,j})} \|g_j\|_\infty \leq \log(c_+/c_-).$$

Step 1.3.2. First, we upper bound the quantity

$$\sup_{g_j \in \mathcal{G}_j(r_{0,j})} |(P_j - P_{n,j})g_j|.$$

For some function class $\mathcal{A}$ and iid Rademacher random variables ζ_t, independent of the data, denote

$$\mathbb{E}_\zeta R_{n,j}\mathcal{A} = \mathbb{E}_\zeta \left[\sup_{h \in \mathcal{A}} \left| \frac{1}{N} \sum_{t=m+1}^{n-m-1} \zeta_t h(X_{t+1,j}, \tilde{Z}_{t,j}) \right| \right].$$

By Theorem 2.1 of Bartlett et al. (2005), for every $\tilde{r} > 0$, with probability at least $1 - 2e^{-\tilde{r}}$, we have

$$\sup_{g_j \in \mathcal{G}_j(r_{0,j})} |(P_j - P_{n,j})g_j| \leq C \left(\mathbb{E}_\zeta R_{n,j}\mathcal{G}_j(r_{0,j}) + r_{0,j} M_j^{\omega_2} \sqrt{\frac{\tilde{r}}{N}} + M_j^{\omega_2} \frac{\tilde{r}}{N} \right), \tag{72}$$

for some constant $C > 0$.

Step 1.3.3. We split this step into two parts.

Step 1.3.3.1. Let $\mathcal{M}_{n,j}^{\mathrm{DNN}}$ be the ReLU DNN function class with the architecture from equation 48. Let $\log \mathcal{T}(u, \mathcal{A}, \|\cdot\|_n)$ denote the metric entropy, and let $\mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})$ denote the pseudo VC dimension of the function class. Observe that, for $u \in (0, 2c_+]$, we have

$$\begin{aligned} \log \mathcal{T}(u, \mathcal{M}_{n,j}^{\mathrm{MDN}}, \|\cdot\|_{n,j}) &\leq 3M_j \log \mathcal{T}\left(\frac{u}{C M_j^{3/2+3\omega_2}}, \mathcal{M}_{n,j}^{\mathrm{DNN}}, \|\cdot\|_{n,j} \right) \\ &\leq C M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}}) \log \left(\frac{C n M_j^{3/2+3\omega_2}}{u} \right), \end{aligned} \tag{73}$$

where the first inequality holds by Lemma 8 (Lemma 4 of Zhou et al. (2023)), and the second inequality follows by Theorems 12.2 and 14.1 in Anthony & Bartlett (1999). Note that the normalization in Assumption 8 is chosen so that the normalization map is Lipschitz, so the same arguments in the proof of Lemma 8 apply with a change in constants.

By Lemma 9, applied to the number of parameters $W_{g,j,\ell}(\epsilon) + U_{g,j,\ell}(\epsilon)$ and the number of hidden layers $H_{g,j,\ell}(\epsilon)$ from equation 48, the pseudo VC dimension of the DNN function class is upper bounded by

$$\mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}}) \leq C M_j^{(k+j-1)/\gamma} \epsilon^{-(k+j-1)/\gamma} (\log(M_j/\epsilon) + 1)^3. \tag{74}$$

Step 1.3.3.2. We justify the comparison of the population norm and empirical norm (which we use in Step 1.3.4) by following the arguments from Appendix A.2.1 of Farrell et al. (2021). For every $r_{0,j} > 0$ and $\tilde{r} > 0$ satisfying

$$r_{0,j} M_j^{\omega_2} \geq c_{\mathrm{loc}} \sqrt{\frac{\max\left(1, M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})\right) \log(n M_j) + \tilde{r}}{N}}, \tag{75}$$

where $c_{\mathrm{loc}} > 0$ is a sufficiently large constant, there exists a constant $c_{\mathrm{cmp}} > 0$, independent of $n, M_j, r_{0,j}, \tilde{r}$, such that, with probability at least $1 - 2e^{-\tilde{r}}$, we have

$$\mathcal{H}_j(r_{0,j}) \subseteq \left\{ h : \|h\|_{n,j} \leq c_{\mathrm{cmp}} r_{0,j} M_j^{\omega_2} \right\}. \tag{76}$$

This class inclusion holds for the following reasons. By equation 71, $\sup_h P_j h^2 \leq c_f^+ (r_{0,j} M_j^{\omega_2})^2$; also $\|h\|_\infty \leq 2c_+$ by the density bounds from Step 1.1. The empirical-Rademacher version of Theorem 2.1 of Bartlett et al. (2005), applied to the class of squared functions, and contraction give, with probability at least $1 - 2e^{-\tilde{r}}$,

$$\sup_{h \in \mathcal{H}_j(r_{0,j})} P_{n,j} h^2 \leq c_f^+ (r_{0,j} M_j^{\omega_2})^2 + C \left(\mathbb{E}_\zeta R_{n,j}\mathcal{H}_j(r_{0,j}) + (r_{0,j} M_j^{\omega_2}) \sqrt{\frac{\tilde{r}}{N}} + \frac{\tilde{r}}{N} \right).$$

Also $\mathrm{Var}(h^2) \leq 4c_+^2 P_j h^2$ and the square function is $4c_+$-Lipschitz on $[-2c_+, 2c_+]$. Dudley's inequality and equation 73 yield

$$\mathbb{E}_\zeta R_{n,j}\mathcal{H}_j(r_{0,j}) \leq \frac{C}{N} + C\sqrt{\sup_{h\in\mathcal{H}_j(r_{0,j})} P_{n,j}h^2}\sqrt{\frac{\max\left(1, M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})\right)\log(nM_j)}{N}}.$$

Combining the previous inequalities and using the inequality $ab \leq \delta a^2 + b^2/(4\delta)$ yields

$$\sup_{h\in\mathcal{H}_j(r_{0,j})} P_{n,j}h^2 \leq C(r_{0,j}M_j^{\omega_2})^2 + C\frac{\max\left(1, M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})\right)\log(nM_j) + \tilde{r}}{N},$$

so equation 75 implies equation 76.

Step 1.3.4. Suppose equation 75 holds. If equation 76 holds, which it does with probability at least $1 - 2e^{-\tilde{r}}$, the contraction and Dudley chaining arguments from Step 1.3 of Zhou et al. (2023) give the following inequalities:

$$\begin{aligned}
&\mathbb{E}_\zeta R_{n,j}\mathcal{G}_j(r_{0,j}) \\
&= \mathbb{E}_\zeta\left[R_{n,j}\left\{g_j : g_j = \log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_j,\right.\right. \\
&\left.\left. f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}}, \left\|f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j} \leq r_{0,j}M_j^{\omega_2}\right\}\right] \\
&\leq \frac{2}{c_-}\mathbb{E}_\zeta\left[R_{n,j}\left\{f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} : f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}},\right.\right. \\
&\left.\left.\left\|f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j} \leq r_{0,j}M_j^{\omega_2}\right\}\right] \\
&\leq \frac{2}{c_-}\mathbb{E}_\zeta R_{n,j}\left\{f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}} : f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}},\right. \\
&\left.\left\|f_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{n,j} \leq c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}\right\} \\
&\leq \frac{2}{c_-}\inf_{0<\alpha<c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}}\left(4\alpha + \frac{12}{\sqrt{N}}\int_\alpha^{c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}} \sqrt{1 + \log\mathcal{T}\left(\delta, \mathcal{M}_{n,j}^{\mathrm{MDN}}, \|\cdot\|_{n,j}\right)}d\delta\right) \\
&\leq \frac{2}{c_-}\inf_{0<\alpha<c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}} \\
&\left(4\alpha + \frac{12}{\sqrt{N}}\int_\alpha^{c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}} \sqrt{1 + 3M_j\log\mathcal{T}\left(\frac{\delta}{CM_j^{3/2+3\omega_2}}, \mathcal{M}_{n,j}^{\mathrm{DNN}}, \|\cdot\|_{n,j}\right)}d\delta\right),
\end{aligned}$$

where the first inequality follows by contraction since the logarithm is c_-^{-1}-Lipschitz, the second is by the localization argument used in Step 1.3 of Zhou et al. (2023) (following Appendix A.2 of Farrell et al. (2021)), i.e., the class inclusion from equation 76, the third is from Dudley's chaining (Mendelson, 2003), and the fourth is from equation 73.

Applying Theorems 12.2 and 14.1 of Anthony & Bartlett (1999) as in Zhou et al. (2023) yields

$$\begin{aligned}
&\mathbb{E}_\zeta R_{n,j}\mathcal{G}_j(r_{0,j}) \\
&\leq \frac{2}{c_-}\inf_{0<\alpha<c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}}\left(4\alpha + \frac{C}{\sqrt{N}}\int_\alpha^{c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}} \sqrt{M_j\mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})\log\left(\frac{CnM_j^{3/2+3\omega_2}}{\delta}\right)}d\delta\right).
\end{aligned}$$

Similar to Zhou et al. (2023), let

$$\alpha = c_{\mathrm{cmp}}r_{0,j}M_j^{\omega_2}\min\left(\frac{1}{2}, \sqrt{\frac{M_j\mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N}}\right),$$

so that, whenever equation 75 holds, there exists a constant $C_4 > 0$, such that, with probability at least $1 - 2e^{-\tilde{r}}$,

$$\mathbb{E}_\zeta R_{n,j}\mathcal{G}_j(r_{0,j}) \leq C_4 r_{0,j} M_j^{\omega_2} \sqrt{\frac{M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N} \log(nM_j)}. \tag{77}$$

Plugging equation 77 into equation 72, we obtain, with probability at least $1 - 6e^{-\tilde{r}}$,

$$\begin{aligned} &\sup_{f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}}(r_{0,j})} \left| (P_j - P_{n,j}) \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_j \right] \right| \\ &\leq C \left(r_{0,j} M_j^{\omega_2} \sqrt{\frac{M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N} \log(nM_j)} + r_{0,j} M_j^{\omega_2} \sqrt{\frac{\tilde{r}}{N}} + M_j^{\omega_2} \frac{\tilde{r}}{N} \right), \end{aligned} \tag{78}$$

for some constant $C > 0$.

Step 2. This step corresponds to Step 2 in the proof of Theorem 3 in Zhou et al. (2023).

Step 2.1. We establish that, for each $j \in [d]$, the process

$$(\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top, \quad t \in \mathbb{Z},$$

is stationary and β-mixing.

Denote the stationary process from Assumption 11 by

$$V_t = (X_t^\top, Y_t^\top, Z_t^\top, \xi_t^\top)^\top, \quad t \in \mathbb{Z}. \tag{79}$$

Each process $(\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top$, $t \in \mathbb{Z}$, is stationary because there exists a Borel map Ψ_j such that

$$(\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top = \Psi_j(V_t, V_{t+1}).$$

For each $j \in [d]$, denote the β-mixing coefficient of $(\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top$, $t \in \mathbb{Z}$, by $\beta_j(i)$, and that of V_t, $t \in \mathbb{Z}$, by $\beta_V(i)$, with a slight abuse of notation. Note that, for every $t' \in \mathbb{Z}$ and $i > 1$,

$$\sigma((\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top : t \leq t') \subseteq \sigma(V_t : t \leq t' + 1),$$

and

$$\sigma((\tilde{Z}_{t,j}^\top, X_{t+1,j})^\top : t \geq t' + i) \subseteq \sigma(V_t : t \geq t' + i),$$

hence

$$\beta_j(i) \leq \beta_V(i-1) \leq c_{\mathrm{mix}}^1 \exp(-c_{\mathrm{mix}}^2 (i-1)), \quad i \geq 2,$$

by the definition of the β-mixing coefficient from Definition 3 because the σ-algebras are nested, where c_{mix}^1 and c_{mix}^2 are from Assumption 11. Therefore, there exist constants $c_{\mathrm{mix}}^{1'}, c_{\mathrm{mix}}^{2'} > 0$, such that

$$\max_{j \in [d]} \beta_j(i) \leq c_{\mathrm{mix}}^{1'} \exp(-c_{\mathrm{mix}}^{2'} i), \quad i \geq 1.$$

Step 2.2. We use Berbee (1979) coupling following Step 2 in the proof of Theorem 3 from Zhou et al. (2023). For some block length $b_n \in \mathbb{N}$, partition the process into blocks of the form

$$U_{i,j} = \left((\tilde{Z}_{m+t,j}^\top, X_{m+t+1,j})^\top \right)_{t=ib_n+1}^{(i+1)b_n}, \quad i \geq 0.$$

By the decoupling lemma, i.e., Lemma 10 (Lemma 5 in Zhou et al. (2023)), there exist copies

$$U_{i,j}^0 = \left(((\tilde{Z}_{m+t,j}^0)^\top, X_{m+t+1,j}^0)^\top \right)_{t=ib_n+1}^{(i+1)b_n},$$

such that each $U_{i,j}^0$ has the same distribution as $U_{i,j}$, the sequences $(U_{2i,j}^0)_{i \geq 0}$ and $(U_{2i+1,j}^0)_{i \geq 0}$ are each iid, and

$$\mathbb{P}(U_{i,j}^0 \neq U_{i,j}) \leq \beta_j(b_n). \tag{80}$$

Using a union bound over the (at most) $\lceil N/b_n \rceil$ blocks that intersect with the first N observations, all of these blocks satisfy $U^0_{i,j} = U_{i,j}$ simultaneously with probability at least

$$1 - \left\lceil \frac{N}{b_n} \right\rceil \beta_j(b_n). \tag{81}$$

For the couplings $(X^0_{t+1,j}, \tilde{Z}^0_{t,j})_t$ introduced previously, define

$$P^0_{n,j} h = \frac{1}{N} \sum_{t=m+1}^{n-m-1} h(X^0_{t+1,j}, \tilde{Z}^0_{t,j}).$$

As in equations (27) and (30) of Zhou et al. (2023), with probability at least that of equation 81,

$$P_{n,j} \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f^{\text{app}}_{j,\epsilon} \right] = P^0_{n,j} \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f^{\text{app}}_{j,\epsilon} \right], \tag{82}$$

$$(P_j - P_{n,j}) \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}^{\text{oracle}}_j \right] = (P_j - P^0_{n,j}) \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}^{\text{oracle}}_j \right]. \tag{83}$$

Next, as in equation (28) of Zhou et al. (2023), we separate the even and odd blocks. For notational simplicity, suppose that $2b_n$ divides N. Then

$$\begin{aligned} P^0_{n,j} h = &\frac{1}{2} \frac{2b_n}{N} \sum_{i=0}^{N/(2b_n)-1} \frac{1}{b_n} \sum_{\ell=1}^{b_n} h\left(X^0_{m+2ib_n+\ell+1,j}, \tilde{Z}^0_{m+2ib_n+\ell,j} \right) \\ &+ \frac{1}{2} \frac{2b_n}{N} \sum_{i=0}^{N/(2b_n)-1} \frac{1}{b_n} \sum_{\ell=1}^{b_n} h\left(X^0_{m+(2i+1)b_n+\ell+1,j}, \tilde{Z}^0_{m+(2i+1)b_n+\ell,j} \right). \end{aligned}$$

Note that, for a fixed measurable function h, each quantity

$$\frac{1}{b_n} \sum_{\ell=1}^{b_n} h\left(X^0_{m+2ib_n+\ell+1,j}, \tilde{Z}^0_{m+2ib_n+\ell,j} \right), \quad \frac{1}{b_n} \sum_{\ell=1}^{b_n} h\left(X^0_{m+(2i+1)b_n+\ell+1,j}, \tilde{Z}^0_{m+(2i+1)b_n+\ell,j} \right),$$

is an iid block average with expectation $P_j h$, so the average consists of two quantities involving iid block averages, each with $N/(2b_n)$ blocks. When $2b_n$ does not divide N, the remaining observations are fewer than $2b_n$, so they contribute $O(b_n/n) = O(\log n/n)$ to the average and thus they do not affect the final rate.

Finally, choose

$$b_n = \left\lceil \frac{2}{c^{2'}_{\text{mix}}} \log n \right\rceil,$$

so that

$$\left\lceil \frac{N}{b_n} \right\rceil \beta_j(b_n) \le c^{1'}_{\text{mix}} \left\lceil \frac{N}{b_n} \right\rceil n^{-2} \le c^{1'}_{\text{mix}} n^{-1}.$$

Therefore, with probability at least $1 - O(n^{-1})$, every block that intersects with the first N observations is equal to its coupled copy. Hence, the coupling identities equation 82 and equation 83 hold simultaneously. Moreover, since $b_n \asymp \log n$, the even-indexed block subsequence, $(U^0_{2i,j})_{i \ge 0}$, and the odd-indexed block subsequence, $(U^0_{2i+1,j})_{i \ge 0}$, each contain $N/(2b_n) \asymp n/\log(n)$ complete blocks in the first N observations.

Step 2.3. Next, as in Zhou et al. (2023), we use the iid sequences of blocks from Step 2.2 to extend the bounds from Steps 1.2.4 and 1.3.4 to the setting of dependent sequences.

In each of the even-indexed and odd-indexed block subsequences, there are $N/(2b_n)$ iid blocks. Therefore, the iid concentration and empirical process arguments may be applied to the even-indexed or odd-indexed block subsequence with sample size $N/(2b_n)$. Also, by Jensen's inequality averaging over observations within each block does not increase the bounds from Step 1, so the covering number bounds remain valid for the corresponding block averages.

For the comparison of the population norm and empirical norm, we apply Step 1.3.3.2 to the iid coupled blocks. By Jensen's inequality, the block averages of h^2 have mean $P_j h^2$ and variance bounded by $4c_+^2 P_j h^2$, so the covering and Dudley bounds continue to hold for the block averages. Since the number of iid blocks is proportional to N/b_n, the corresponding condition is

$$r_{0,j} M_j^{\omega_2} \geq c_{\mathrm{loc}} \sqrt{\frac{b_n}{N} \left(\max\left(1, M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})\right) \log(nM_j) + \tilde{r}\right)}. \tag{84}$$

The following inequalities are analogous to equations (29) and (31) in Zhou et al. (2023). Applying the argument from Step 1.2.4 to the even-indexed and odd-indexed iid sequences of blocks and combining the two bounds yields, for every $\tilde{r} > 0$, with probability at least $1 - Ke^{-\tilde{r}}$ for some constant $K > 0$,

$$\begin{aligned} &P_{n,j}^0 \left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_{j,\epsilon}^{\mathrm{app}}\right] \\ &\leq C\left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)^2 \\ &\quad + C\left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)\sqrt{\frac{b_n\tilde{r}}{N}} + C\left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)\frac{b_n\tilde{r}}{N}. \end{aligned} \tag{85}$$

Similarly, applying the argument from Step 1.3 to each iid block subsequence, provided equation 84 holds, yields, with probability at least $1 - K'e^{-\tilde{r}}$ for some constant $K' > 0$,

$$\begin{aligned} &\sup_{f_j \in \mathcal{M}_{n,j}^{\mathrm{MDN}}(r_{0,j})} \left|(P_j - P_{n,j}^0)\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log f_j\right]\right| \\ &\leq C\left(r_{0,j} M_j^{\omega_2}\sqrt{\frac{b_n M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N}\log(nM_j)} + r_{0,j}M_j^{\omega_2}\sqrt{\frac{b_n\tilde{r}}{N}} + M_j^{\omega_2}\frac{b_n\tilde{r}}{N}\right). \end{aligned} \tag{86}$$

Step 2.4. Next, we obtain the convergence rate for each conditional density estimator by combining the previous bounds. Using the bounds on $K_j(\hat{f}_j^{\mathrm{oracle}})$ from Step 1.1, taking $f_j^{\mathrm{DNN}} = f_{j,\epsilon}^{\mathrm{app}}$, the coupling identities from Step 2.2, equation 85 and equation 86, we have that, for every $\tilde{r} > 0$ satisfying equation 84, with probability at least $1 - K''e^{-\tilde{r}} - O(n^{-1})$ for some constant $K'' > 0$,

$$\begin{aligned} \left\|\hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j}^2 \leq C\Bigg(&r_{0,j}M_j^{\omega_2}\sqrt{\frac{b_n M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N}\log(nM_j)} \\ &+ r_{0,j}M_j^{\omega_2}\sqrt{\frac{b_n\tilde{r}}{N}} + M_j^{\omega_2}\frac{b_n\tilde{r}}{N} \\ &+ \left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)^2 + \left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)\sqrt{\frac{b_n\tilde{r}}{N}} \\ &+ \left(M_j^{4\omega_2}\epsilon + M_j^{-\omega_1}\right)\frac{b_n\tilde{r}}{N}\Bigg), \end{aligned}$$

for some constant $C > 0$.

As in Step 2 of Zhou et al. (2023), we apply the arguments from Appendices A.2.1–A.2.4 of Farrell et al. (2021) to recursively improve the upper bound $r_{0,j}$. Since $b_n \asymp \log n$, this yields, for every $\tilde{r} > 0$, with probability at least $1 - K'''e^{-\tilde{r}} - O(n^{-1})$ for some constant $K''' > 0$,

$$\begin{aligned} \left\|\hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j} \leq C\Bigg(&M_j^{4\omega_2}\left[\sqrt{\frac{M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N}}\log(nM_j)\right. \\ &\left. + \log n\sqrt{\frac{\log\log n + \tilde{r}}{N}} + \epsilon\log n\right] + M_j^{-\omega_1}\Bigg). \end{aligned}$$

By equation 74,

$$\mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}}) \leq C M_j^{(k+j-1)/\gamma}\epsilon^{-(k+j-1)/\gamma}\left(\log(M_j/\epsilon) + 1\right)^3.$$

Recall from Step 1.2.2 that

$$\epsilon = n^{-\gamma/(2\gamma+k+j-1)}.$$

Therefore, we have

$$\sqrt{\frac{M_j \mathrm{Pdim}(\mathcal{M}_{n,j}^{\mathrm{DNN}})}{N}} \log(nM_j) \le C M_j^{\frac{\gamma+k+j-1}{2\gamma}} n^{-\frac{\gamma}{2\gamma+k+j-1}} \log^3(nM_j).$$

Also, setting $\tilde{r} = \log n$, we have

$$\log n \sqrt{\frac{\log\log n + \tilde{r}}{N}} \le C n^{-1/2} \log^{3/2} n, \quad \epsilon \log n = n^{-\frac{\gamma}{2\gamma+k+j-1}} \log n.$$

Since $M_j \ge 1$ and $\gamma/(2\gamma + k + j - 1) < 1/2$, both terms are bounded by

$$C M_j^{\frac{\gamma+k+j-1}{2\gamma}} n^{-\frac{\gamma}{2\gamma+k+j-1}} \log^3(nM_j).$$

Putting it all together, for each $j \in [d]$, with probability at least $1 - O(n^{-1})$, we have

$$\left\| \hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j} \le C \left(M_j^{-\omega_1} + M_j^{\frac{\gamma+k+j-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+j-1}} \log^3(nM_j) \right), \quad (87)$$

for some constant $C > 0$.

Step 3. Lastly, we combine the conditional density bounds for each $j \in [d]$ from Step 2.4 to obtain the bound for the conditional joint density. By the factorization of $f_{X_{t+1}|Z_t}$ from equation 23, the definition of $\hat{f}^{\mathrm{oracle}}$ in equation 35, and a telescoping identity for the difference of two products,

$$\begin{aligned}
&\hat{f}^{\mathrm{oracle}}(x \mid z) - f_{X_{t+1}|Z_t}(x \mid z) \\
&= \prod_{j=1}^{d} \hat{f}_j^{\mathrm{oracle}}(x_j \mid \tilde{z}_j) - \prod_{j=1}^{d} f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \\
&= \sum_{j=1}^{d} \left(\prod_{i=1}^{j-1} f_{X_{t+1,i}|\tilde{Z}_{t,i}}(x_i \mid \tilde{z}_i) \right) \left(\hat{f}_j^{\mathrm{oracle}}(x_j \mid \tilde{z}_j) - f_{X_{t+1,j}|\tilde{Z}_{t,j}}(x_j \mid \tilde{z}_j) \right) \\
&\left(\prod_{i=j+1}^{d} \hat{f}_i^{\mathrm{oracle}}(x_i \mid \tilde{z}_i) \right).
\end{aligned}$$

For every $j \in [d]$,

$$\begin{aligned}
&\left\| \left(\prod_{i=1}^{j-1} f_{X_{t+1,i}|\tilde{Z}_{t,i}} \right) \left(\hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right) \left(\prod_{i=j+1}^{d} \hat{f}_i^{\mathrm{oracle}} \right) \right\|_{L^2}^2 \\
&\le (c_f^+)^{j-1} c_{\mathrm{MDN}}^{2(d-j)} \left(\prod_{i=j+1}^{d} \lambda_{\mathcal{X}_i}(\mathcal{X}_i) \right) \left\| \hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2 \\
&\le C \left\| \hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}^2,
\end{aligned}$$

for some constant $C > 0$, where the first inequality holds because the true conditional density factors are upper bounded by c_f^+ and the estimated MDN factors are upper bounded by c_{MDN} by Assumptions 7 and 8, and the second inequality follows because d is fixed and $\mathcal{X}$ is compact by Assumption 3. Hence, by the triangle inequality, we have

$$\left\| \hat{f}^{\mathrm{oracle}} - f_{X_{t+1}|Z_t} \right\|_{L^2} \le C \sum_{j=1}^{d} \left\| \hat{f}_j^{\mathrm{oracle}} - f_{X_{t+1,j}|\tilde{Z}_{t,j}} \right\|_{L^2,j}. \quad (88)$$

Since d is fixed, a union bound over equation 87 implies that, with probability at least $1 - O(n^{-1})$, the bounds hold simultaneously for each $j \in [d]$.

Recall $M_- = \min_{j\in[d]}(M_j)$ and $M_+ = \max_{j\in[d]}(M_j)$. Observe that $M_j^{-\omega_1} \leq M_-^{-\omega_1}$, and $1 \leq M_j \leq M_+$, so we have

$$M_j^{\frac{\gamma+k+j-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+j-1}} \log^3(nM_j) \leq M_+^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM_+).$$

Combining these inequalities with equation 88 yields, with probability at least $1 - O(n^{-1})$,

$$\left\| \hat{f}^{\text{oracle}} - f_{X_{t+1}|Z_t} \right\|_{L^2} \leq c \left(M_-^{-\omega_1} + M_+^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM_+) \right),$$

for some constant $c > 0$, which proves equation 36. □

D.2 Proof of Theorem 7

The proof is broken into two main steps.

Step 1. First, we prove via induction that, for every $L \in \mathbb{N}$, we have

$$X_{t+1:t+L} \perp\!\!\!\perp X_{-\infty:t} \mid S_t.$$

Step 1.1. The base case $L = 1$, that

$$X_{t+1} \perp\!\!\!\perp X_{-\infty:t} \mid S_t, \tag{89}$$

holds by the one-step-ahead predictive sufficiency property from equation 39.

Step 1.2. Suppose that, for some $L \in \mathbb{N}$, we have

$$X_{t+1:t+L} \perp\!\!\!\perp X_{-\infty:t} \mid S_t. \tag{90}$$

Equivalently, by Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)),

$$\mathcal{L}(X_{t+1:t+L} \mid \mathcal{F}_t) = \mathcal{L}(X_{t+1:t+L} \mid S_t), \tag{91}$$

almost surely.

Observe that, for each $L \in \mathbb{N}$, there exists a measurable function $\tau^{(L)}$ that is defined as the recursive application of τ to S_t and $X_{t+1:t+L}$, such that

$$S_{t+L} = \tau^{(L)}(S_t, X_{t+1:t+L}) = \tau(\cdots \tau(\tau(S_t, X_{t+1}), X_{t+2}) \cdots, X_{t+L}). \tag{92}$$

For example, for $L = 1$ and $L = 2$, we have

$$S_{t+1} = \tau^{(1)}(S_t, X_{t+1}) = \tau(S_t, X_{t+1}),$$
$$S_{t+2} = \tau^{(2)}(S_t, X_{t+1:t+2}) = \tau(\tau(S_t, X_{t+1}), X_{t+2}) = \tau(S_{t+1}, X_{t+2}).$$

For any sets $A \in \mathfrak{X}^{\otimes L}$ and $B \in \mathfrak{X}$, where $\mathfrak{X}^{\otimes L}$ is the product σ-algebra, we have, almost surely,

$$\mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid \mathcal{F}_t\right) = \mathbb{E}\left[\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid \mathcal{F}_{t+L}\right) \mid \mathcal{F}_t\right],$$

by the tower property of conditional expectation, and

$$\mathbb{P}\left(X_{t+L+1} \in B \mid \mathcal{F}_{t+L}\right) = \mathbb{P}\left(X_{t+L+1} \in B \mid S_{t+L}\right),$$

by the one-step-ahead predictive sufficiency property from equation 39. Therefore, we have

$$\mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid \mathcal{F}_t\right) = \mathbb{E}\left[\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid S_{t+L}\right) \mid \mathcal{F}_t\right],$$

almost surely, by applying the previous equalities.

By equation 92, the quantity $\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid S_{t+L}\right)$ is a measurable function of $(S_t, X_{t+1:t+L})$, so by equation 91, we have

$$\mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid \mathcal{F}_t\right) = \mathbb{E}\left[\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid S_{t+L}\right) \mid S_t\right],$$

almost surely. On the other hand, by the tower property of conditional expectation and the one-step-ahead predictive sufficiency property from equation 39, we have

$$\begin{aligned}\mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid S_t\right) &= \mathbb{E}\left[\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid \mathcal{F}_{t+L}\right) \mid S_t\right] \\ &= \mathbb{E}\left[\mathbf{1}\{X_{t+1:t+L} \in A\}\mathbb{P}\left(X_{t+L+1} \in B \mid S_{t+L}\right) \mid S_t\right],\end{aligned}$$

almost surely. Therefore, we have, almost surely, that

$$\mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid \mathcal{F}_t\right) = \mathbb{P}\left(X_{t+1:t+L} \in A, X_{t+L+1} \in B \mid S_t\right). \tag{93}$$

The Cartesian products $A \times B$, with $A \in \mathfrak{X}^{\otimes L}$, $B \in \mathfrak{X}$, form a π-system, i.e., it is closed under finite intersections, that generates $\mathfrak{X}^{\otimes(L+1)}$. Moreover, the collection of sets $C \in \mathfrak{X}^{\otimes(L+1)}$ for which the almost sure equality from equation 93 holds is a λ-system, i.e., it is closed under proper differences and increasing limits, and it contains $\mathcal{X}^{L+1}$. Therefore, by the monotone class theorem (Theorem 1.1 in Kallenberg (2021)), the almost sure equality from equation 93 holds for every set $C \in \mathfrak{X}^{\otimes(L+1)}$. Since $\sigma(S_t) \subseteq \mathcal{F}_t$, Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)) implies that

$$X_{t+1:t+L+1} \perp\!\!\!\perp X_{-\infty:t} \mid S_t.$$

Hence, if the inductive hypothesis from equation 90 holds for L, then it holds for $L+1$.

Step 1.3. By equation 39, the inductive hypothesis holds for $L = 1$, so for every $L \in \mathbb{N}$ we have,

$$X_{t+1:t+L} \perp\!\!\!\perp X_{-\infty:t} \mid S_t. \tag{94}$$

Since $S_t = R(X_{-\infty:t})$ by equation 37, we have $\sigma(S_t) \subseteq \mathcal{F}_t$, so by Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)), equation 94 is equivalent to

$$\mathcal{L}(X_{t+1:t+L} \mid \mathcal{F}_t) = \mathcal{L}(X_{t+1:t+L} \mid S_t), \tag{95}$$

almost surely.

Step 2. Second, we extend this statement to the infinite future,

$$X_{t+1:\infty} \perp\!\!\!\perp X_{-\infty:t} \mid S_t,$$

via the monotone class theorem (Theorem 1.1 in Kallenberg (2021)). Define

$$\mathcal{D} = \{A \in \sigma(X_{t+1:\infty}) : \mathbb{P}(A \mid \mathcal{F}_t) = \mathbb{P}(A \mid S_t) \text{ almost surely}\}. \tag{96}$$

We claim that $\mathcal{D}$ is a λ-system, i.e., it is closed under proper differences and increasing limits, and it contains Ω. First, if $A, B \in \mathcal{D}$ with $B \subseteq A$, then by the linearity property of conditional expectation,

$$\mathbb{P}(A \setminus B \mid \mathcal{F}_t) = \mathbb{P}(A \mid \mathcal{F}_t) - \mathbb{P}(B \mid \mathcal{F}_t) = \mathbb{P}(A \mid S_t) - \mathbb{P}(B \mid S_t) = \mathbb{P}(A \setminus B \mid S_t),$$

almost surely, so $A \setminus B \in \mathcal{D}$. Second, if $A_n \in \mathcal{D}$ with $A_n \uparrow A$, then, almost surely,

$$\mathbb{P}(A \mid \mathcal{F}_t) = \lim_{n\to\infty} \mathbb{P}(A_n \mid \mathcal{F}_t) = \lim_{n\to\infty} \mathbb{P}(A_n \mid S_t) = \mathbb{P}(A \mid S_t),$$

by the monotone convergence property of conditional expectation. Third, observe that

$$\mathbb{P}(\Omega \mid \mathcal{F}_t) = \mathbb{P}(\Omega \mid S_t) = 1,$$

almost surely, hence $\Omega \in \mathcal{D}$. Therefore, $\mathcal{D}$ is a λ-system.

Next, we claim that the class

$$\mathcal{C} = \bigcup_{L\in\mathbb{N}} \sigma(X_{t+1:t+L}),$$

is a π-system, i.e., it is closed under finite intersections. If $A, B \in \mathcal{C}$ then there must exist $L_1, L_2 \in \mathbb{N}$ such that $A \in \sigma(X_{t+1:t+L_1})$ and $B \in \sigma(X_{t+1:t+L_2})$, therefore $A \cap B \in \sigma(X_{t+1:t+\max(L_1,L_2)})$ because the sequence of σ-algebras $\sigma(X_{t+1:t+L})$, $L \in \mathbb{N}$, is increasing.

By equation 95, we have that

$$\mathcal{C} \subseteq \mathcal{D}.$$

Thus, by the monotone class theorem (Theorem 1.1 in Kallenberg (2021)), we have that

$$\sigma(\mathcal{C}) \subseteq \mathcal{D}.$$

Moreover, observe that

$$\sigma(\mathcal{C}) = \sigma(X_{t+1:\infty}),$$

so $\sigma(X_{t+1:\infty}) \subseteq \mathcal{D}$. By the definition of $\mathcal{D}$ from equation 96, we have $\mathcal{D} \subseteq \sigma(X_{t+1:\infty})$. Therefore,

$$\mathcal{D} = \sigma(X_{t+1:\infty}).$$

Hence, for every $A \in \sigma(X_{t+1:\infty})$, we have, almost surely, that

$$\mathbb{P}(A \mid \mathcal{F}_t) = \mathbb{P}(A \mid S_t). \tag{97}$$

Since $S_t = R(X_{-\infty:t})$ by equation 37, we have $\sigma(S_t) \subseteq \mathcal{F}_t$. By Doob's conditional independence property (Theorem 8.9 in Kallenberg (2021)), the equality from equation 97 is equivalent to

$$X_{t+1:\infty} \perp\!\!\!\perp X_{-\infty:t} \mid S_t,$$

which is the desired result. □

D.3 Proof of Lemma 1

We may always decompose Y_{t+1} from equation 11 as

$$Y_{t+1} = Z_t + \xi_{t+1}, \tag{98}$$

where $Z_t = \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]$ is from equation 13, and the noise process ξ_{t+1} satisfies

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[Y_{t+1} - Z_t \mid \mathcal{F}_t] = \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] - \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = 0,$$

almost surely, where the first equality is by the definition of $\xi_{t+1} = Y_{t+1} - Z_t$ by rearranging equation 98, and the second equality is by the definition of $Z_t = \mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]$, the linearity property of conditional expectation, and because $\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]$ is $\mathcal{F}_t$-measurable. Next, observe that

$$\begin{aligned}
Y_{t+1} &\overset{(1)}{=} Z_t + \xi_{t+1} \\
&\overset{(2)}{=} \Phi(S_t) + \xi_{t+1} \\
&\overset{(3)}{=} \Phi(\tau^{(m+1)}(S_{t-m-1}, X_{t-m:t})) + \xi_{t+1} \\
&\overset{(4)}{=} \Phi(\tau^{(m+1)}(\Phi^{-1}(Z_{t-m-1}), X_{t-m:t})) + \xi_{t+1} \\
&\overset{(5)}{=} \Phi(\tau^{(m+1)}(\Phi^{-1}(Y_{t-m} - \xi_{t-m}), X_{t-m:t})) + \xi_{t+1} \\
&\overset{(6)}{=} g^*(Y_{t-m}, \xi_{t-m}, X_{t-m:t}) + \xi_{t+1},
\end{aligned} \tag{99}$$

where (1) is from equation 98, (2) is by equation 13, (3) is by defining the function $\tau^{(m+1)}$ as the recursive application of τ from equation 2 a total of $m+1$ times, (4) is because the inverse Φ^{-1} exists since $\Phi : \mathcal{S} \to \Phi(\mathcal{S})$ is bijective by the injectivity of Φ from Assumption 2, (5) is by equation 98 with t replaced by $t - m - 1$, and (6) is by the definition of $g^*(\cdot)$. Specifically, for domain

$$\mathcal{D} = \{(y, e, x) \in [-1, 1]^k \times [-2, 2]^k \times \mathcal{X}^{m+1} : y - e \in \Phi(\mathcal{S}), \tau^{(m+1)}(\Phi^{-1}(y - e), x) \in \mathcal{S}\},$$

where we define the function $g^* : \mathcal{D} \to \Phi(\mathcal{S})$ as

$$g^*(y, e, x) = \Phi(\tau^{(m+1)}(\Phi^{-1}(y - e), x)).$$

□

E Proofs of Main Results

E.1 Proof of Theorem 1

The result follows directly from Theorem 7 since $\mathcal{X}$ and $\mathcal{S}$ are Borel subsets of Euclidean spaces, each equipped with its Borel σ-algebra, hence they are standard Borel spaces.

□

E.2 Proof of Theorem 2

The proof is similar to that of Theorem 1 in Wieck-Sosa & Shalizi (2026a) and Theorem 2 in Wieck-Sosa & Shalizi (2026b). However, their motivations are very different from ours. They consider simulation-based parameter estimation for parametric models of time series, and they estimate *unconditional expectations* of random features *using time-averages* of the observed random features in order to injectively map the parameter space into feature space. In contrast, we estimate *conditional expectations* of random features given the past observations *using neural networks* in order to injectively map the predictive states of a stochastic process into feature space.

The key ingredient is the finite witness theorem (Lemma 3). Lemma 3 can be applied with any $k \geq 2p+1$, not necessarily $k = 2p+1$; see the proof of Proposition 3.6 in Amir et al. (2023).

We begin by showing that the assumptions of Lemma 3 are satisfied.

Denote the collection of parameters for the i-th random Fourier feature, φ_i, by

$$\vartheta_i = (\Omega_{i,1}^\top, \ldots, \Omega_{i,m+1}^\top, \alpha_i)^\top, \tag{100}$$

and let $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$, which is an $\mathbb{R}^{(d(m+1)+1)\times k}$-valued random variable. Denote

$$\mathbb{W} = \mathbb{R}^{d(m+1)} \times (-\pi, \pi).$$

By Assumption 1, the predictive state space $\mathcal{S}$ is a compact subset of $\mathbb{R}^r$ with Hausdorff dimension $\dim_H(\mathcal{S}) = p$. The predictive state space $\mathcal{S}$ is a semialgebraic subset of $\mathbb{R}^r$ by Assumption 1, hence it is σ-subanalytic; see Figure 3.

By construction, the random parameters ϑ_i, $i = 1, \ldots, k$ from equation 100 for the random Fourier features φ_i, $i = 1, \ldots, k$ from equation 10 are each random variables taking values in $\mathbb{W}$, which is semialgebraic, so, as explained in Figure 3, it is σ-subanalytic.

By Assumption 1, the map $(s, \omega) \mapsto \psi_s(\omega)$ admits a jointly real analytic extension to $O \times \mathbb{R}^{(m+1)\times d}$, and hence each $(s, \vartheta_i) \mapsto \int_{\mathcal{X}^{m+1}} \varphi_i(x) dP_s(x)$, $i = 1, \ldots, k$ admits a jointly real analytic extension to $O \times \mathbb{W}$ by using the definition of the characteristic function and each random Fourier feature. Note that the graph of this extension is σ-subanalytic, and that $\mathcal{S}$ is a σ-subanalytic set as discussed previously. Therefore, since σ-subanalytic sets are closed under finite intersections and Cartesian products by Lemma 2, we have that the intersection of the graph of this extension with $\mathcal{S} \times \mathbb{W} \times \mathbb{R}$ is σ-subanalytic.

$\mathbb{W}$ is open and connected, and $(s, \vartheta_i) \mapsto \int_{\mathcal{X}^{m+1}} \varphi_i(x) dP_s(x)$ is analytic as a function of ϑ_i for all fixed values of the predictive state $s \in \mathcal{S}$. Hence, the condition from equation 33 from the finite witness theorem (Lemma 3) is satisfied.

Let $x_1, \ldots, x_{m+1}$ be vectors in $\mathcal{X}$, and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathcal{X}^{m+1}$. Letting $a = (a_1^\top, \ldots, a_{m+1}^\top)^\top \in \mathbb{R}^{d(m+1)}$, $b \in (-\pi, \pi)$, and $c = (a, b) \in \mathbb{W}$, define

$$\phi_c(x) = \cos\left(\sum_{j=1}^{m+1} a_j \cdot x_j + b\right).$$

Define

$$\mathcal{N} = \{(s_1, s_2) \in \mathcal{S}^2 \mid \int_{\mathcal{X}^{m+1}} \phi_c(x) dP_{s_1}(x) = \int_{\mathcal{X}^{m+1}} \phi_c(x) dP_{s_2}(x),\ \forall c \in \mathbb{W}\},$$

as the set of pairs of predictive state values that yield identical conditional expectations of the functions ϕ_c for all values of $c \in \mathbb{W}$. By the finite witness theorem (Lemma 3), for generic parameters $(\vartheta_1, \ldots, \vartheta_k) \in \mathbb{W}^k$ of the random Fourier features $(\varphi_1, \ldots, \varphi_k)$, we have the set equality

$$\mathcal{N} = \{(s_1, s_2) \in \mathcal{S}^2 \mid \int_{\mathcal{X}^{m+1}} \varphi_i(x) dP_{s_1}(x) = \int_{\mathcal{X}^{m+1}} \varphi_i(x) dP_{s_2}(x),\ \forall i = 1, \ldots, k\}.$$

By the discussion above Lemma 3, generic here means Π-almost every ϑ, where Π is the joint distribution from equation 100 which is absolutely continuous with respect to the Lebesgue measure.

Denoting the $\mathbb{R}^{(m+1)\times d}$-valued random matrix by $\Omega_i = (\Omega_{i,1}, \ldots, \Omega_{i,m+1})^\top$, we have

$$\begin{aligned}
&\int_{\mathcal{X}^{m+1}} \varphi_i(x) dP_s(x) \\
&= \int_{\mathcal{X}^{m+1}} \cos\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j + \alpha_i\right) dP_s(x) \\
&= \cos(\alpha_i) \int_{\mathcal{X}^{m+1}} \cos\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j\right) dP_s(x) \\
&- \sin(\alpha_i) \int_{\mathcal{X}^{m+1}} \sin\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j\right) dP_s(x) \\
&= \cos(\alpha_i)\, \mathrm{Re}(\psi_s(\Omega_i)) - \sin(\alpha_i)\, \mathrm{Im}(\psi_s(\Omega_i)) \\
&= M_{\vartheta_i}(s),
\end{aligned}$$

where ψ_s is the characteristic function. If the pair of predictive state values $(s_1, s_2) \in \mathcal{N}$, then $M_{\vartheta_i}(s_1) = M_{\vartheta_i}(s_2)$ for every possible value of $\vartheta_i = (\mathrm{Vec}(\Omega_i), \alpha_i) \in \mathbb{W}$. In particular, the conditional expectation values of the random Fourier features correspond to the real and imaginary parts of the characteristic function by taking $\alpha_i = 0$ and $\alpha_i = -\pi/2$, respectively, so by the uniqueness theorem (Lemma 4) the corresponding conditional distributions must be equal, $P_{s_1} = P_{s_2}$, if $(s_1, s_2) \in \mathcal{N}$.

However, by Assumption 1, the map from the predictive state value to the conditional distribution, $s \mapsto P_s$, is one-to-one, i.e., if $s_1 \neq s_2$ then $P_{s_1} \neq P_{s_2}$, so we must have

$$(s_1, s_2) \in \mathcal{N} \implies (s_1, s_2) \in \{(s_1, s_2) \in \mathcal{S}^2 : s_1 = s_2\} = \{(s, s) : s \in \mathcal{S}\},$$

so $\mathcal{N} \subseteq \{(s, s) : s \in \mathcal{S}\}$. Conversely, we have $\{(s, s) : s \in \mathcal{S}\} \subseteq \mathcal{N}$ trivially. Therefore,

$$\mathcal{N} = \{(s, s) : s \in \mathcal{S}\}. \tag{101}$$

Putting it all together, we have: for Π-almost any random parameters $(\vartheta_1, \ldots, \vartheta_k)$ for the random Fourier features $(\varphi_1, \ldots, \varphi_k)$, for every pair of predictive state values $s_1, s_2 \in \mathcal{S}$, if $s_1 \neq s_2$, then the conditional expectations of the random Fourier features are not equal, $\Phi(s_1) \neq \Phi(s_2)$. That is, the map $s \mapsto \Phi(s)$ is one-to-one for generic choices of $k \geq 2p + 1$ random Fourier features.

□

E.3 Proof of Theorem 3

First, we show that $\sigma(Z_t) \subseteq \sigma(S_t)$ by showing that Z_t can be written as a measurable function of S_t. This holds because $Z_t = \Phi(S_t)$ by definition from equation 13, and Φ is measurable by the joint real analyticity of the characteristic function from Assumption 1.

Second, we show that $\sigma(S_t) \subseteq \sigma(Z_t)$ by showing that S_t can be written as a measurable function of Z_t. Observe that the inverse Φ^{-1} exists since $\Phi : \mathcal{S} \to \Phi(\mathcal{S})$ is bijective by the injectivity of Φ from Assumption 2. Also, observe that the inverse Φ^{-1} is measurable since it is continuous by Lemma 5, which applies because (1) Φ is one-to-one by Assumption 2, (2) Φ is continuous by the joint real analyticity of the characteristic function from Assumption 1, and (3) $\mathcal{S}$ is compact by Assumption 1. Therefore, $S_t = \Phi^{-1}(Z_t)$ and Φ^{-1} is measurable, hence $\sigma(S_t) \subseteq \sigma(Z_t)$.

Putting it all together, we have the desired result that

$$Q_{Z_t} = \mathcal{L}(X_{t+1} \mid Z_t) = \mathcal{L}(X_{t+1} \mid S_t),$$

almost surely, where the first equality is by definition and the second is because $\sigma(S_t) = \sigma(Z_t)$.

□

E.4 Proof of Theorem 4

In this proof, we derive the convergence rates for the estimator $\hat{f}$ from equation 26 based on the learned predictive state representations.

Denote the sample size by

$$N = n - 2m - 1,$$

where n is the number of observations of X_t, $t = 1, \ldots, n$, and m is from the random Fourier features from equation 10.

We break the proof into three steps.

Step 1. First, we obtain the average RNN estimation error.

For each fixed choice of RNN parameters, the initialization and recurrence from equation 20 imply that the prediction $h_t = \lambda(H_{t+1})$ is $\mathcal{F}_t$-measurable, and that

$$Z_t \in \arg\min_h \mathbb{E} \left\| Y_{t+1} - h \right\|_2^2,$$

where the argmin is over all $\mathcal{F}_t$-measurable random vectors. This follows by the definition of ξ_{t+1} from equation 16 and expanding the square,

$$\begin{aligned} \left\| Y_{t+1} - h_t \right\|_2^2 &= \left\| Z_t - h_t + \xi_{t+1} \right\|_2^2 \\ &= \left\| Z_t - h_t \right\|_2^2 + \left\| \xi_{t+1} \right\|_2^2 + 2\xi_{t+1}^\top (Z_t - h_t), \end{aligned}$$

so

$$\mathbb{E}\left[\left\| Y_{t+1} - h_t \right\|_2^2 - \left\| \xi_{t+1} \right\|_2^2 \right] = \mathbb{E} \left\| Z_t - h_t \right\|_2^2 + 2\mathbb{E}\left[\xi_{t+1}^\top (Z_t - h_t) \right],$$

where the cross term is zero because

$$\mathbb{E}\left[\xi_{t+1}^\top (Z_t - h_t) \right] = \mathbb{E}\left[\mathbb{E}\left[\xi_{t+1}^\top (Z_t - h_t) \mid \mathcal{F}_t \right] \right] = \mathbb{E}\left[(Z_t - h_t)^\top \mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] \right] = 0. \tag{102}$$

Hence, we have

$$\mathbb{E} \left\| Y_{t+1} - h_t \right\|_2^2 = \mathbb{E} \left\| \xi_{t+1} \right\|_2^2 + \mathbb{E} \left\| h_t - Z_t \right\|_2^2.$$

The first term on the right is independent of h_t, so Z_t is the minimizer of $\mathbb{E} \left\| Y_{t+1} - h_t \right\|_2^2$ over all $\mathcal{F}_t$-measurable random vectors h_t.

The target $Y_{t+1} = \varphi(X_{t+1:t+m+1})$ need not be $\mathcal{F}_{t+1}$-measurable when $m > 0$. However, the RNN input $I_{t+1} = (X_{t-m:t}, Y_{t-m})$ is $\mathcal{F}_t$-measurable since $Y_{t-m} = \varphi(X_{t-m:t})$ so the cross term from equation 102 is zero for every RNN predictor $h_t = \lambda(H_{t+1})$, which establishes the conditional orthogonality condition used in the concentration arguments in the proofs of Theorems 1 and 2 in Chen et al. (2026).

The concentration and covering arguments used in the proofs of Theorems 1 and 2 in Chen et al. (2026) require certain boundedness and mixing conditions. As we will show, these conditions are implied by our Assumptions 3 and 11, respectively.

Next, we check that these boundedness and mixing conditions hold, so that the Bernstein inequality (Lemma 6 in Chen et al. (2026)) used in Lemma 8 in Chen et al. (2026) is applicable.

First, we check that the boundedness conditions hold. For fixed network parameters and an integer $\ell \geq 0$, let $h_t^{(\ell)}$ be the prediction obtained by replacing the hidden state immediately before the most recent $\ell + 1$ recurrent updates by zero, as in Definition 3 in the supplementary material of Chen et al. (2026). Since $h_t^{(\ell)}$ is $\mathcal{F}_t$-measurable, we have

$$\mathbb{E}\left[\xi_{t+1}^\top (Z_t - h_t^{(\ell)}) \right] = 0,$$

by the same arguments as equation 102. Also, observe that

$$\mathbb{E}\left| \xi_{t+1}^\top (Z_t - h_t^{(\ell)}) \right|^2 \leq 4k\mathbb{E} \left\| Z_t - h_t^{(\ell)} \right\|_2^2,$$

by Cauchy-Schwarz and since $\|\xi_{t+1}\|_2^2 \le 4k$, and that

$$\left|\xi_{t+1}^\top (Z_t - h_t^{(\ell)})\right| \le \|\xi_{t+1}\|_1 \left\|Z_t - h_t^{(\ell)}\right\|_\infty \le 4k,$$

almost surely, by Hölder's inequality and since $\|\xi_{t+1}\|_\infty \le 2$ and $\left\|Z_t - h_t^{(\ell)}\right\|_\infty \le 2$.

Second, we check that the mixing condition holds. Recall $V_t = (X_t^\top, Y_t^\top, Z_t^\top, \xi_t^\top)^\top$ from equation 79. The random variable $\xi_{t+1}^\top (Z_t - h_t^{(\ell)})$ is measurable with respect to $\sigma(V_{t-\ell-m}, \ldots, V_{t+1})$, so the β-mixing coefficient at lag $a > \ell + m + 1$ of the process $(\xi_{t+1}^\top [Z_t - h_t^{(\ell)}])_{t\in\mathbb{Z}}$ is upper bounded by $\beta_V(a - \ell - m - 1)$. Therefore, the argument involving the interlaced subsequences from Lemma 8 in the supplementary material of Chen et al. (2026) applies with spacing $\ell + m + 2$ and truncation $\ell \asymp w_{h,n} \log^2 n$. For the subsequences obtained by taking every $(\ell+m+2)$-th term, the mixing coefficient at lag $a \ge 1$ is bounded by

$$\beta_V(a(\ell + m + 2) - \ell - m - 1) = \beta_V((a-1)(\ell + m + 2) + 1) \le c_{\text{mix}}^1 \exp(-c_{\text{mix}}^2 a).$$

Note that we assume β-mixing, whereas Chen et al. (2026) only assume α-mixing. Since α-mixing coefficients are upper bounded by β-mixing coefficients (see, e.g., Bradley (2005)), the concentration inequality used in Lemma 8 of the supplementary material of Chen et al. (2026) applies.

Choose $\ell = \lceil C^* w_{h,n} \log^2 n \rceil$ for a sufficiently large constant $C^* > 0$. Lemma 7 from the supplementary material of Chen et al. (2026) controls the difference between the truncated and full RNN predictions. The covering and union bounding arguments from Lemma 10 in Chen et al. (2026), yields a concentration bound that holds uniformly over the RNN class, so in particular it applies to the fitted RNN.

Therefore, analogous arguments from the proof of Theorem 1 in Chen et al. (2026) apply with the RNN input $I_{t+1} = (X_{t-m:t}, Y_{t-m})$ and the filtration $\mathcal{F}_t$. Similarly, analogous RNN approximation arguments from the proof of Theorem 2 in Chen et al. (2026) apply to the representation

$$Z_t = g^*(Y_{t-m}, \xi_{t-m}, X_{t-m:t}),$$

from Lemma 1, where the input dimension of $g^*(\cdot)$ is $2k + d(m+1)$. Since m from equation 10 is fixed, using these lags only changes constants in the proof of Theorem 2 in Chen et al. (2026) due to changes in the width of the recurrent layer.

By equation 18 in the proof of Theorem 1 in Chen et al. (2026), Theorem 2 of Chen et al. (2026) with the widths and depths from the statement of Theorem 4, and Markov's inequality, we have

$$\frac{1}{N}\sum_{t=m+1}^{n-m-1} \left\|\hat{Z}_t - Z_t\right\|_2^2 = O_{\mathbb{P}}\left(n^{-\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11}\right). \tag{103}$$

Step 2. Second, we obtain the MDN estimation error.

Recall $P_j, P_{n,j}, K_j$, and $\|\cdot\|_{L^2,j}$ from the proof of Theorem 6. Since the estimator from equation 26 is an argmax, by Assumption 8 and Jensen's inequality, we have

$$\begin{aligned} P_{n,j}[\log g_{n,j} - \log \hat{f}_j] &\le \frac{2\Lambda_n}{N}\sum_{t=m+1}^{n-m-1} \left\|\hat{Z}_t - Z_t\right\|_2^\kappa \\ &\le 2\Lambda_n \left(\frac{1}{N}\sum_{t=m+1}^{n-m-1} \left\|\hat{Z}_t - Z_t\right\|_2^2\right)^{\kappa/2}, \end{aligned} \tag{104}$$

where $P_{n,j}$ is from equation 41 and $g_{n,j}$ is the approximation from Assumption 10. Note that the factor of 2 accounts for replacing the covariates in both $g_{n,j}$ and $\hat{f}_j$. By equation 44, equation 27, and equation 28, we have

$$K_j(g_{n,j}) \le C a_{n,j}^2.$$

Note that the mean value theorem applied to $\log$ and the density bounds give a log-likelihood ratio bounded by $C a_{n,j}$. Hence, the argument from equation 85, with $\tilde{r} = 3 \log n$ and the coupling identities from equation 82 and equation 83, imply that, with probability at least $1 - O(n^{-1})$,

$$P_{n,j}(\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log g_{n,j}) \le C(M^{-\omega_1} + M^{\frac{\gamma+k+j-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+j-1}} \log^3(nM))^2. \tag{105}$$

Observe that

$$\begin{aligned}\frac{c_f^-}{2c_+^2}\left\|\hat{f}_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j}^2 \le (P_j - P_{n,j})\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}_j\right] \\ + P_{n,j}\left[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log g_{n,j}\right] \\ + 2\Lambda_n\left(\frac{1}{N}\sum_{t=m+1}^{n-m-1}\left\|\hat{Z}_t - Z_t\right\|_2^2\right)^{\kappa/2},\end{aligned} \tag{106}$$

by using the lower bound from equation 45, decomposing $K_j(\hat{f}_j)$ from equation 42 into

$$\begin{aligned}K_j(\hat{f}_j) = (P_j - P_{n,j})[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log \hat{f}_j] \\ + P_{n,j}[\log f_{X_{t+1,j}|\tilde{Z}_{t,j}} - \log g_{n,j}] \\ + P_{n,j}[\log g_{n,j} - \log \hat{f}_j],\end{aligned}$$

and applying the inequality from equation 104 to the last term.

Applying the inequality from equation 105 to the second term in equation 106, and applying equation 86 and equation 74 to the first term in equation 106 and then applying the same recursive improvement argument used in Step 2.4 of the proof of Theorem 6, yields, with probability $1-O(n^{-1})$,

$$\begin{aligned}\frac{c_f^-}{4c_+^2}\left\|\hat{f}_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j}^2 \le C\left(M^{-\omega_1} + M^{\frac{\gamma+k+j-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+j-1}}\log^3(nM)\right)^2 \\ + 2\Lambda_n\left(\frac{1}{N}\sum_{t=m+1}^{n-m-1}\left\|\hat{Z}_t - Z_t\right\|_2^2\right)^{\kappa/2}.\end{aligned} \tag{107}$$

Step 3. Third, we obtain the convergence rate.

The factorization from equation 23 and the argument from equation 88 give

$$\left\|\hat{f} - f_{X_{t+1}|Z_t}\right\|_{L^2} \le C\sum_{j=1}^{d}\left\|\hat{f}_j - f_{X_{t+1,j}|\tilde{Z}_{t,j}}\right\|_{L^2,j}.$$

Taking square roots in equation 107, using the inequality $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, substituting the RNN estimation error from equation 103, and applying Markov's inequality gives

$$\left\|\hat{f} - f_{X_{t+1}|Z_t}\right\|_{L^2} = O_{\mathbb{P}}\begin{pmatrix} M^{-\omega_1} + M^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}}\log^3(nM) \\ + \Lambda_n^{1/2} n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)}(\log n)^{11\kappa/4}\end{pmatrix}.$$

Note that the MDN with $j = d$ has the slowest convergence rate due to the greater number of covariates, so the upper bound is largest at $j = d$. □

E.5 Proof of Theorem 5

We break the proof into three steps.

Step 1. Choose $g_{n,j}$ from Assumption 10, and let

$$g_n(x \mid z) = \prod_{j=1}^{d} g_{n,j}(x_j \mid \tilde{z}_j). \tag{108}$$

For our estimator $\hat{f}$ from equation 26 and g_n from equation 108, define

$$q_n(z) = \left\|\hat{f}(\cdot \mid z) - g_n(\cdot \mid z)\right\|_{L^2(\mathcal{X})}. \tag{109}$$

By equation 27, the factorization from equation 23, the density bounds in Assumptions 7 and 8, and the compactness from Assumption 3, we have

$$\sup_{z\in\mathcal{Z}} \left\| g_n(\cdot \mid z) - f_{X_{t+1}|Z_t}(\cdot \mid z) \right\|_{L^2(\mathcal{X})} \le C \sum_{j=1}^{d} a_{n,j} \le C \left(M^{-\omega_1} + M^{4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \right), \quad (110)$$

where the last inequality uses the facts that $M_j = M$, $j \le d$, and d is fixed.

By the triangle inequality, equation 110, and equation 29, we have

$$\begin{aligned} \|q_n\|_{L^2(P_{Z_t})} &\le \left\| \hat{f} - f_{X_{t+1}|Z_t} \right\|_{L^2} + \left\| g_n - f_{X_{t+1}|Z_t} \right\|_{L^2} \\ &= O_{\mathbb{P}} \begin{pmatrix} M^{-\omega_1} + M^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM) \\ + \Lambda_n^{1/2} n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/4} \end{pmatrix}. \end{aligned} \quad (111)$$

For either $f = \hat{f}$ or $f = g_n$, the Hölder condition in Assumption 8 implies that

$$\sup_{x\in\mathcal{X}} |\log f(x \mid z) - \log f(x \mid z')| \le \Lambda_n d \left\| z - z' \right\|_2^{\kappa}, \quad z, z' \in [-1,1]^k. \quad (112)$$

Moreover, by the mean value theorem applied to $\exp(\cdot)$, i.e., $|e^a - e^b| \le e^{\max(a,b)}|a-b|$, for all $x \in \mathcal{X}$ and $z, z' \in \mathcal{Z}$, we have

$$\begin{aligned} |f(x \mid z) - f(x \mid z')| &= |\exp(\log f(x \mid z)) - \exp(\log f(x \mid z'))| \\ &\le \exp(\max(\log f(x \mid z), \log f(x \mid z')))| \log f(x \mid z) - \log f(x \mid z')| \\ &\le \exp(d \log c_{\mathrm{MDN}})| \log f(x \mid z) - \log f(x \mid z')| \\ &\le c_{\mathrm{MDN}}^d | \log f(x \mid z) - \log f(x \mid z')| \\ &\le c_{\mathrm{MDN}}^d \Lambda_n d \left\| z - z' \right\|_2^{\kappa}, \end{aligned}$$

because each of the d MDN factors is upper bounded by c_{MDN} from Assumption 8. Define

$$C_0 = \max\left(1, 2d c_{\mathrm{MDN}}^d \left(\int_{\mathcal{X}} 1 dx \right)^{1/2} \right), \quad (113)$$

which is finite due to the compactness of $\mathcal{X}$ from Assumption 3. For $q_n(\cdot)$ from equation 109, for $z \in \mathcal{Z}$, the triangle inequality and Assumption 8 yield

$$q_n(z) \le \left(\int_{\mathcal{X}} \left(|\hat{f}(x \mid z)| + |g_n(x \mid z)| \right)^2 dx \right)^{1/2} \le 2 c_{\mathrm{MDN}}^d \left(\int_{\mathcal{X}} 1 dx \right)^{1/2} \le C_0, \quad (114)$$

where the last inequality holds by equation 113 and because $d \ge 1$. By the reverse triangle inequality, the triangle inequality, and the inequalities above, for $z, z' \in \mathcal{Z}$, we have

$$\begin{aligned} |q_n(z) - q_n(z')| &\le \left\| \hat{f}(\cdot \mid z) - \hat{f}(\cdot \mid z') \right\|_{L^2(\mathcal{X})} + \| g_n(\cdot \mid z) - g_n(\cdot \mid z') \|_{L^2(\mathcal{X})} \\ &\le 2 d c_{\mathrm{MDN}}^d \left(\int_{\mathcal{X}} 1 dx \right)^{1/2} \Lambda_n \left\| z - z' \right\|_2^{\kappa} \\ &\le C_0 \Lambda_n \left\| z - z' \right\|_2^{\kappa}. \end{aligned} \quad (115)$$

Step 2. Second, we evaluate $q_n(\cdot)$ from equation 109 at the true predictive state embedding.

Let $\delta > 0$ and $0 < v < C_0$. Divide each of the k coordinate intervals $[-1,1]$ from $[-1,1]^k$ into

$$\left\lceil 2\sqrt{k} \left(\frac{2 C_0 \Lambda_n}{v} \right)^{1/\kappa} \right\rceil, \quad (116)$$

consecutive intervals of equal length. To make the intervals disjoint, we take each interval to include its left endpoint but not its right endpoint, except for the final interval, which also includes 1.

Denote the Cartesian products of these intervals across the k coordinates by $B_1, \ldots, B_{J'}$ for some $J' \in \mathbb{N}$, each of which is a subset of $[-1, 1]^k$. Define

$$A_i = \mathcal{Z} \cap B_i,$$

and denote the non-empty sets, i.e., such that $\mathcal{Z} \cap B_i \neq \varnothing$, as $A_1, \ldots, A_J$ for some $J \in \mathbb{N}$. Note that, by construction, the sets $A_1, \ldots, A_J$ are disjoint, measurable subsets of $\mathcal{Z}$ that satisfy

$$\mathcal{Z} = \bigcup_{i=1}^{J} A_i.$$

Moreover, since each coordinate interval has length at most $k^{-1/2}(v/(2C_0\Lambda_n))^{1/\kappa}$, we have

$$\operatorname{diam}(A_i) = \sup_{a_1, a_2 \in A_i} \|a_1 - a_2\|_2 \leq \left(\frac{v}{2C_0\Lambda_n}\right)^{1/\kappa}.$$

By equation 116, the number of intervals in each of the k coordinates is at most a constant multiple of

$$\left(\frac{\Lambda_n}{v}\right)^{1/\kappa},$$

so, since k is fixed, we have

$$J \leq C\left(\frac{\Lambda_n}{v}\right)^{k/\kappa}, \tag{117}$$

for some constant $C > 0$. Therefore, by equation 115 and the construction above, for any $z, z' \in A_i$,

$$|q_n(z) - q_n(z')| \leq C_0\Lambda_n \operatorname{diam}(A_i)^\kappa \leq \frac{v}{2}. \tag{118}$$

Suppose $q_n(Z_{n+L-1}) > v$ and $\|q_n\|_{L^2(P_{Z_t})} \leq \delta$. In this case, if $Z_{n+L-1} \in A_i$, then,

$$q_n(z) \geq q_n(Z_{n+L-1}) - v/2 > v/2,$$

for every $z \in A_i$. Hence, we have

$$\frac{v^2}{4} P_{Z_t}(A_i) \leq \int_{A_i} q_n(z)^2 dP_{Z_t}(z) \leq \delta^2, \tag{119}$$

where the second inequality is because $\|q_n\|_{L^2(P_{Z_t})} \leq \delta$, and the first inequality follows from equation 118 because when $q_n(Z_{n+L-1}) > v$, if $Z_{n+L-1} \in A_i$, then,

$$q_n(z) \geq q_n(Z_{n+L-1}) - |q_n(z) - q_n(Z_{n+L-1})| > v - \frac{v}{2} = \frac{v}{2},$$

for every $z \in A_i$. Multiplying equation 119 by $4/v^2$ gives us

$$P_{Z_t}(A_i) \leq \frac{4}{v^2}\delta^2. \tag{120}$$

By Assumption 11, the process is stationary, so $\mathbb{P}(Z_{n+L-1} \in A_i) = P_{Z_t}(A_i)$. Therefore, we have

$$\begin{aligned}
\mathbb{P}(q_n(Z_{n+L-1}) > v) &\leq \mathbb{P}\left(q_n(Z_{n+L-1}) > v, \|q_n\|_{L^2(P_{Z_t})} > \delta\right) \\
&\quad + \mathbb{P}\left(q_n(Z_{n+L-1}) > v, \|q_n\|_{L^2(P_{Z_t})} \leq \delta\right) \\
&\leq \mathbb{P}\left(\|q_n\|_{L^2(P_{Z_t})} > \delta\right) + \sum_{i: P_{Z_t}(A_i) \leq 4\delta^2/v^2} P_{Z_t}(A_i) \\
&\leq \mathbb{P}\left(\|q_n\|_{L^2(P_{Z_t})} > \delta\right) + C\delta^2\Lambda_n^{k/\kappa} v^{-2-k/\kappa},
\end{aligned} \tag{121}$$

where the first inequality is by set inclusion, the last inequality is by equation 117 and equation 120, and the second inequality is by set inclusion and the following argument: if the event in the second

term, $\{q_n(Z_{n+L-1}) > v, \|q_n\|_{L^2(P_{Z_t})} \leq \delta\}$, occurs, then for some i such that $P_{Z_t}(A_i) \leq \frac{4}{v^2}\delta^2$ we have $Z_{n+L-1} \in A_i$ by equation 120, so a union bound gives the inequality.

Recall C_0 from equation 113. For some $K_0 > 0$, substituting $v = K_0 \Lambda_n^{k/(2\kappa+k)} \delta^{2\kappa/(2\kappa+k)}$ into equation 121 yields

$$\mathbb{P}\left(q_n(Z_{n+L-1}) > K_0 \Lambda_n^{\frac{k}{2\kappa+k}} \delta^{\frac{2\kappa}{2\kappa+k}}\right) \leq \mathbb{P}\left(\|q_n\|_{L^2(P_{Z_t})} > \delta\right) + C K_0^{-2-k/\kappa}. \tag{122}$$

Note that, for $v \geq C_0$, $\mathbb{P}(q_n(Z_{n+L-1}) > v)$ from equation 121 is zero by equation 114.

For any given $\epsilon > 0$, choose δ to be a sufficiently large constant multiple of the expression inside the $O_{\mathbb{P}}(\cdot)$ from equation 111, which makes the first term on the right side of equation 122 at most $\epsilon/2$ for large n. Choosing the constant K_0 so that $C K_0^{-2-k/\kappa}$ from equation 122 is less than or equal to $\epsilon/2$ implies that

$$\begin{aligned} &q_n(Z_{n+L-1}) \\ &= O_{\mathbb{P}}\left(\Lambda_n^{\frac{k}{2\kappa+k}} \begin{bmatrix} M^{-\omega_1} + M^{\frac{\gamma+k+d-1}{2\gamma}+4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}} \log^3(nM) \\ + \Lambda_n^{1/2} n^{-\frac{\kappa}{4}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/4} \end{bmatrix}^{\frac{2\kappa}{2\kappa+k}}\right). \end{aligned} \tag{123}$$

Step 3. Third, we substitute the estimated predictive state.

By Assumption 6 and equation 19, we have

$$\hat{Z}_{n+L-1} \in [-1, 1]^k.$$

Also, by equation 30 and Markov's inequality, we have

$$\begin{aligned} &\Lambda_n d \left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^{\kappa} \\ &= O_{\mathbb{P}}\left(\Lambda_n n^{-\frac{\kappa}{2}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/2}\right) \\ &= o_{\mathbb{P}}(1), \end{aligned} \tag{124}$$

which tends to zero by the assumption in the theorem statement because this rate from equation 124 is the square of the last summand of the expression inside $O_{\mathbb{P}}(\cdot)$ from equation 29.

For $x \in \mathcal{X}$, we have

$$\begin{aligned} &|\hat{f}(x \mid \hat{Z}_{n+L-1}) - \hat{f}(x \mid Z_{n+L-1})| \\ &= \hat{f}(x \mid Z_{n+L-1}) \left|\exp\left(\log \hat{f}(x \mid \hat{Z}_{n+L-1}) - \log \hat{f}(x \mid Z_{n+L-1})\right) - 1\right| \\ &\leq \hat{f}(x \mid Z_{n+L-1}) \left[\exp\left(\Lambda_n d \left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^{\kappa}\right) - 1\right] \\ &\leq c_{\text{MDN}}^d \left[\exp\left(\Lambda_n d \left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^{\kappa}\right) - 1\right], \end{aligned}$$

by properties of $\log(\cdot)$ and $\exp(\cdot)$, factoring out $\hat{f}(x \mid Z_{n+L-1})$, equation 112, and Assumption 8.

By equation 124,

$$\Lambda_n d \left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^{\kappa} = o_{\mathbb{P}}(1),$$

which implies that this term can be upper bounded by one with probability tending to one. By the mean value theorem, with probability tending to one, we have

$$\begin{aligned} &\left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - \hat{f}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &\leq 3 c_{\text{MDN}}^d \left(\int_{\mathcal{X}} 1 dx\right)^{1/2} \Lambda_n d \left\|\hat{Z}_{n+L-1} - Z_{n+L-1}\right\|_2^{\kappa}. \end{aligned}$$

Since $\mathcal{X}$ is compact by Assumption 3, applying equation 124 yields

$$\begin{aligned}&\left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - \hat{f}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &= O_{\mathbb{P}}\left(\Lambda_n n^{-\frac{\kappa}{2}\min\left(\frac{\eta}{\eta+4k+2d(m+1)}, \frac{2k+d(m+1)+3}{10k+5d(m+1)+3}\right)} (\log n)^{11\kappa/2}\right).\end{aligned} \tag{125}$$

Since the process is stationary by Assumption 11, (X_{n+L}, Z_{n+L-1}) has the same distribution as (X_{t+1}, Z_t), so we can use the same (version of) the conditional density. Observe that

$$\begin{aligned}&\left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - f_{X_{n+L}|Z_{n+L-1}}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &\leq \left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - \hat{f}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &+ \left\|\hat{f}(\cdot \mid Z_{n+L-1}) - g_n(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &+ \left\|g_n(\cdot \mid Z_{n+L-1}) - f_{X_{t+1}|Z_t}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})},\end{aligned}$$

by adding and subtracting both $\hat{f}(\cdot \mid Z_{n+L-1})$ and $g_n(\cdot \mid Z_{n+L-1})$, and by applying the triangle inequality. By the definition of $q_n(\cdot)$ from equation 109,

$$q_n(Z_{n+L-1}) = \left\|\hat{f}(\cdot \mid Z_{n+L-1}) - g_n(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})}.$$

Also, since $Z_{n+L-1} \in \mathcal{Z}$ almost surely, equation 110 yields

$$\begin{aligned}&\left\|g_n(\cdot \mid Z_{n+L-1}) - f_{X_{t+1}|Z_t}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &\leq \sup_{z\in\mathcal{Z}} \left\|g_n(\cdot \mid z) - f_{X_{t+1}|Z_t}(\cdot \mid z)\right\|_{L^2(\mathcal{X})} \\ &\leq C\left(M^{-\omega_1} + M^{4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}}\right).\end{aligned}$$

Combining the above inequalities, we have

$$\begin{aligned}&\left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - f_{X_{n+L}|Z_{n+L-1}}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &\leq \left\|\hat{f}(\cdot \mid \hat{Z}_{n+L-1}) - \hat{f}(\cdot \mid Z_{n+L-1})\right\|_{L^2(\mathcal{X})} \\ &+ q_n(Z_{n+L-1}) \\ &+ C\left(M^{-\omega_1} + M^{4\omega_2} n^{-\frac{\gamma}{2\gamma+k+d-1}}\right).\end{aligned} \tag{126}$$

The first term has the rate from equation 125, which is bounded by the square of the expression in the square brackets from equation 123. The second term has the rate from equation 123. The final term is bounded by a constant times the same expression in the square brackets from equation 123.

Hence, each term in equation 126 is bounded by the rate in equation 31 because the expression in the square brackets from equation 123 tends to zero, because $\Lambda_n \geq 1$ $2\kappa/(2\kappa+k) \in (0,1)$, and because $u^2 \leq u \leq \Lambda_n^{\frac{k}{2\kappa+k}} u^{\frac{2\kappa}{2\kappa+k}}$ for $0 < u \leq 1$. This establishes the desired rate from equation 31.

□