PNAS
RESEARCH ARTICLE | STATISTICS

Estimating dynamic models by matching random features

Michael Wieck-Sosa[a] and Cosma Rohilla Shalizi[a,1]

This manuscript was compiled on July 17, 2026

Scientists increasingly express their ideas as dynamic models of complex processes. It is often much easier to simulate these models than to calculate the probability of their generating a particular outcome, making likelihood-based estimation infeasible. Existing likelihood-free approaches rely either on manually chosen summary statistics or on representations learned by neural networks. The former is error-prone and laborious, while the latter is computationally intensive, leaving many scientists in a difficult position. We show that the parameters of many dynamical models can be estimated by matching a small number of random features of the observed and simulated data. Specifically, we adapt results from nonlinear dynamics to show that models with a p-dimensional parameter can generically be identified from just $2p+1$ random features. This leads to a simple and general estimation strategy based on matching random features, which converges on the true parameter values under mild regularity conditions. We apply the methodology to discrete-time dynamical systems and differential equations with observational noise used in ecology and epidemiology. More broadly, our results serve as the foundation for a class of random feature methods for simulation-based estimation and inference.

simulation-based inference | parameter estimation | time series | nonlinear dynamics | embedding

Across scientific disciplines, dynamic models are used to describe the mechanisms that generate natural phenomena over time. Yet the same mechanistic detail that makes these models scientifically compelling often places conventional likelihood-based statistical inference out of reach. In such cases, simulation offers a way forward. Simulation-based methods assess parameter values θ by comparing certain characteristics of the observations with those of the synthetic data generated from the model at θ. However, it is often unclear which characteristics should be compared and which should be treated as noise.

Many simulation-based methods rely on carefully crafted summary statistics. These summary statistics are designed to be sensitive to changes in the parameter values and to ensure identifiability. Typically, the summary statistics are user-chosen (1, 2) or learned (3, 4).

However, both approaches face difficulties. Manually selecting summary statistics requires substantial problem-specific effort for scientists. On the other hand, neural network-based methods are computationally demanding and sensitive to implementation choices.

We propose to replace both hand-crafted and learned summary statistics with *randomly chosen* functions of the data. We establish that models with a p-dimensional parameter θ can be identified with generic choices of just $2p+1$ random features. Our results suggest that simulation-based methods can be made far more automatic through random features, without meaningfully reducing precision.

Our approach combines classical probabilistic and statistical ideas with modern random feature methods from machine learning (5) and embedding results from the "geometry from a time series" literature (6). It is well-established that probability distributions can be uniquely characterized by the expectations they assign a sufficiently rich set of test functions. Indeed, probabilists *define* convergence of distributions as the convergence of the

Significance

Throughout history, scientists have created models to understand dynamic phenomena in the natural world, such as ecosystems and infectious diseases. However, the detailed mechanisms that make these models useful also make them difficult to fit because the likelihood function is typically unavailable. We introduce a broadly applicable likelihood-free parameter estimation method based on matching a small number of randomly chosen features of observed and simulated data. In contrast with existing likelihood-free approaches, we avoid hand-crafted summary statistics and costly neural-network training. Our method provides a path to streamlined simulation-based parameter estimation for complex scientific models.

Author affiliations: [a]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA 15213

M.W.-S. and C.S. both contributed to the conceptualization, writing, and editing of all versions of this manuscript.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: cshalizi@cmu.edu

expectations of all bounded, continuous functions; see Chapter 4 of (7). The space of all test functions is too rich to be practically useful, but it is well-understood that there are smaller, more practical function spaces that are still rich enough to be convergence-determining. For distributions on Euclidean spaces, the expectations of Fourier basis functions are convergence-determining; see Chapter 5 of (7).

Our first innovation is to realize that a *parametric* model can only produce a very limited variety of distributions. Using ideas from embedding theory, we show that the expectations of "almost any" $2p+1$ smooth test functions will uniquely characterize the p-dimensional manifold of distributions produced by a model with a p-dimensional parameter space. Our second innovation is to realize that it is enough to pick $2p+1$ functions *at random* from one of those sufficiently-rich, convergence-determining classes. This can be done without looking at the data, or even understanding much about the model. The expectations of $2p+1$ randomly chosen functions will typically characterize the p-dimensional parameter of the model, and the mapping from expectation values to parameter values will be smooth. Figure 1B illustrates this in a simple example: independent and identically distributed (iid) observations from a Gaussian distribution with unknown mean.

This, in and of itself, is a result at the level of probability distributions and expectations, rather than data and estimators; it is about identification, rather than consistency. To get an actual parameter estimation procedure, we need to make additional assumptions which allow us to reliably approximate expectation values from observations. We show this is possible in two regimes: long-time (or ergodic) asymptotics, and dense-sampling (or infill) asymptotics.

The first regime applies when there is a stationary distribution, or at least an asymptotically stationary distribution. Because a stationary process is, by definition, one where the distribution of values over time intervals is invariant under translation, we can meaningfully speak of *the* expectation value of any function. The expectation values of the $2p+1$ random features can then be reliably approximated simply by taking averages of those features over time, using classical ergodic theory and mixing theory. Figure 1A illustrates the key steps.

The second regime applies to nonstationary processes, precisely those where the distribution changes meaningfully over time, and so the expectations of random features also vary over time. The correct framework for such situations is that of so-called "infill asymptotics", where we get more and more observations related to each local structure of the underlying nonstationary process. In this regime, the local expectations of the $2p+1$ random features can be reliably approximated by averages of those features over rolling windows, which can be specified without knowledge of the data generating process.

In both regimes, we use mathematical representations of stochastic processes as transformations of sequences of noise inputs. Most dynamic models admit such a representation. This can be checked by recursively expressing each observation in terms of its past noise inputs. Examples include discrete-time dynamical systems and ordinary differential equations (ODEs) with observational noise, discretized stochastic differential equations (SDEs),

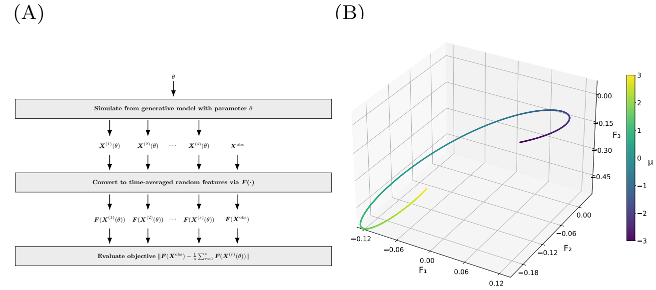

Fig. 1. Key steps of the time-average procedure and example with the normal distribution.
(A) We observe a time series $\mathbf{X}^{\text{obs}}$ and simulate s replicate time series $\mathbf{X}^{(r)}(\theta)$, $r = 1, \ldots, s$, from the model, given some value for the p-dimensional parameter θ. Each time series is converted to time-averages of the same $k = 2p+1$ random features via $\mathbf{F}(\cdot)$, which yields $\mathbf{F}(\mathbf{X}^{\text{obs}})$ and $\mathbf{F}(\mathbf{X}^{(r)}(\theta))$, $r = 1, \ldots, s$. We then evaluate the objective by calculating the Euclidean distance $\left\|\mathbf{F}(\mathbf{X}^{\text{obs}}) - \frac{1}{s}\sum_{r=1}^{s} \mathbf{F}(\mathbf{X}^{(r)}(\theta))\right\|$. The estimate $\hat{\theta}$ is obtained by minimizing this objective via optimization. The rolling-window procedure for nonstationary time series is similar, but it minimizes distances between rolling-window averages. See Methods for details.
(B) Let $\mathbf{X}^{(r)} = (X_t^{(r)})_{t=1}^{n}$, where $X_t^{(r)} \overset{\text{iid}}{\sim} N(\mu, 1)$, for $t = 1, \ldots, n \equiv 1{,}000$, $r = 1, \ldots, s \equiv 10$. The plot shows $\frac{1}{s}\sum_{r=1}^{s} \mathbf{F}(\mathbf{X}^{(r)}(\theta))$ as a function of the mean parameter $\theta = \mu$ of the normal distribution, where color denotes the value of $\mu \in [-3, 3]$. Three random features identify each μ value, the parameter-to-feature map is smooth, and it has a smooth inverse.

linear and nonlinear time series models, and state-space models.

To establish the convergence of averages to the corresponding expectations, we require some mild, technical regularity assumptions, which say that the impact of any single noise input must eventually decay over time. Under certain conditions, our assumptions are equivalent to mixing-type assumptions concerning the decay of temporal dependence. We also require that the distributions change smoothly with the parameter. Under these assumptions, we provide guarantees that our parameter estimators are "consistent" in the sense of statistical theory, so that our estimates converge to the true parameter values.

Applications to biological dynamic models

We apply our random feature methods to fit prototypical ecological and epidemiological dynamic models. We consider additional data generating processes in the Supplementary Information. Each density plot is constructed from 1,000 independent estimates of the p-dimensional parameter using $2p+1$ random Fourier features (see Methods) and 10 simulations per parameter value. To be clear, each parameter estimate is based on an independent realization of the process and an independent draw of $2p+1$ random features.

The experiments reported below were conducted in Python. We optimized the objective functions using the differential evolution solver `differential_evolution` from the `scipy.optimize` module in the SciPy library, although our experiments indicate that other optimizers perform comparably. For the examples with differential equations (ODEs), we used the Runge–Kutta 5(4) solver `RK45` from the `scipy.integrate` module in SciPy.

Logistic map model. (8) popularized the logistic map as a population model capable of producing periodic and chaotic dynamics. For $t = 1, \ldots, n$, define the latent state by

$$Z_t = \rho Z_{t-1}(1 - Z_{t-1}),$$

where the unknown initial value Z_0 is sampled from a $U[0, 1]$ distribution. We observe

$$X_t = Z_t + \sigma \varepsilon_t,$$

where each $\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1)$. Our goal is to estimate

$$\rho_0 = 3.9, \quad \sigma_0 = 0.1.$$

The logistic map is in a chaotic regime when $\rho = 3.9$; see (9) for more details.

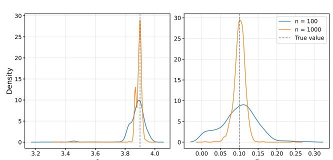

Fig. 2. Logistic map model. Density of parameter estimates using $2p + 1 = 5$ random features.

SIR model. (10) introduced the susceptible-infected-recovered (SIR) model to describe the dynamics of infectious disease transmission. For time horizon $T = 50$, let $(S(v), I(v), R(v))_{v \in [0,T]}$ be defined by the system of ODEs

$$\frac{dS}{dv}(v) = -\frac{\beta}{N} S(v)\, I(v),$$
$$\frac{dI}{dv}(v) = \frac{\beta}{N} S(v)\, I(v) - \gamma I(v),$$
$$\frac{dR}{dv}(v) = \gamma I(v),$$

where $S(v)/N$, $I(v)/N$, and $R(v)/N$ denote the fractions of the population that are susceptible, infected, and recovered at time $v \in [0, T]$, respectively, $\beta > 0$ is the transmission rate, and $\gamma > 0$ is the recovery rate. Let the known initial values be $S(0) = 980{,}000$, $I(0) = 20{,}000$, $R(0) = 0$, where the number of individuals is $N = S(0) + I(0) + R(0) = 1{,}000{,}000$.

For $t = 1, \ldots, n$, we randomly test $\mathrm{TE}_t = \lfloor N/100 \rfloor$ people using a test with sensitivity $\mathrm{SE}_t = 0.95$ and specificity $\mathrm{SP}_t = 0.99$. We observe the proportion of positive tests

$$X_t = Z_t / \mathrm{TE}_t,$$
$$Z_t \sim \mathrm{Binomial}\left(\mathrm{TE}_t, \pi_t\right),$$
$$\pi_t = \mathrm{SE}_t I(Tt/n)/N + (1 - \mathrm{SP}_t)\left(1 - I(Tt/n)/N\right).$$

This model allows TE_t, SE_t, and SP_t to change over time t, although we keep them constant for simplicity. We aim to estimate

$$\beta_0 = 0.5, \quad \gamma_0 = 0.2.$$

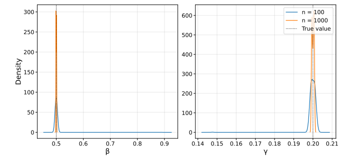

Fig. 3. SIR model. Density of parameter estimates using $2p + 1 = 5$ random features.

Lotka–Volterra model. (11, 12) introduced a model to describe predator–prey population dynamics. For time horizon $T = 20$, let $(Z^{(1)}(v), Z^{(2)}(v))_{v \in [0,T]}$ be defined by the system of ODEs

$$\frac{dZ^{(1)}}{dv}(v) = \alpha Z^{(1)}(v) - \beta Z^{(1)}(v) Z^{(2)}(v),$$
$$\frac{dZ^{(2)}}{dv}(v) = \delta Z^{(1)}(v) Z^{(2)}(v) - \gamma Z^{(2)}(v),$$

where $Z^{(1)}(v)$ and $Z^{(2)}(v)$ represent the prey and predator population densities (i.e. the number of individuals per unit area) at time $v \in [0, T]$, respectively. The known initial values are $Z^{(1)}(0) = 21$ and $Z^{(2)}(0) = 9$. The components of the parameter are the prey growth rate $\alpha > 0$, predation rate $\beta > 0$, predator growth rate $\delta > 0$, and predator death rate $\gamma > 0$. For $t = 1, \ldots, n$, we observe

$$X_t^{(1)} = Z^{(1)}(Tt/n) + \sigma^{(1)} \varepsilon_t^{(1)},$$
$$X_t^{(2)} = Z^{(2)}(Tt/n) + \sigma^{(2)} \varepsilon_t^{(2)},$$

where each $\varepsilon_t^{(1)}, \varepsilon_t^{(2)} \overset{\text{iid}}{\sim} N(0, 1)$ and $\sigma^{(1)}$, $\sigma^{(2)}$ control the noise level. We aim to estimate

$$\alpha_0 = 0.8, \quad \beta_0 = 0.08, \quad \delta_0 = 0.06,$$
$$\gamma_0 = 1.2, \quad \sigma_0^{(1)} = 0.13, \quad \sigma_0^{(2)} = 0.19.$$

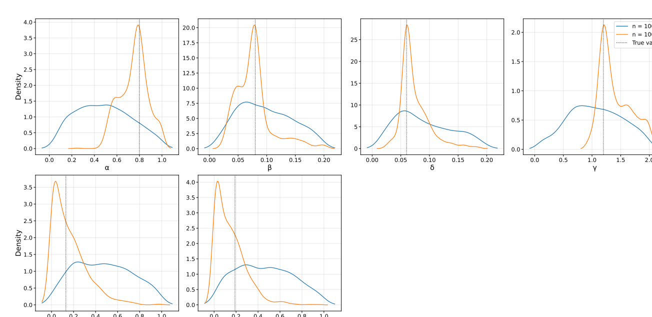

Fig. 4. Lotka–Volterra model. Density of parameter estimates using $2p + 1 = 13$ random features.

Discussion

We present an alternative to the existing paradigms for simulation-based estimation. Typically, simulation-based methods rely on user-chosen summary statistics or on neural network-based learned representations. Both approaches have large costs. In contrast, we reduce the problem of estimating a p-dimensional parameter to the problem of matching $2p + 1$ *random* features of the observed and simulated data. We propose two random feature estimators, and we prove that they are consistent for asymptotically stationary and generally nonstationary dynamic models, respectively.

The principle of matching random features is applicable beyond dynamic models for time series data. This suggests a broader research program on random feature methods for various data types, including spatiotemporal, network, point process, and functional data. The key technical ingredients are: concentration inequalities under dependence, control of the variation of the data-generating mechanism (e.g., over space and time), and a suitable class of random features for the dependence structure.

Random features may also be used for uncertainty quantification. For example, measures of uncertainty may be obtained using a model-based bootstrap procedure. That is, one first estimates the parameter, then simulates datasets from the fitted model and re-estimates the parameter based on each simulated dataset, then finally calculates the empirical standard deviation or variance-covariance matrix of the re-estimates. Additionally, in a separate manuscript, we establish the asymptotic normality of the random feature estimators under a slightly stronger set of assumptions, i.e. after appropriate rescaling, their deviations from the true parameter values converge to a mean-zero Gaussian distribution. Simulation results concerning uncertainty quantification will be reported in that manuscript.

The Supplementary Information contains the assumptions, theoretical results, proofs, and technical lemmata. It also includes additional simulation results. Further, a selective review of related literature is provided.

Methods

In this section, we introduce the notation, setting, random features, and estimators.

Notation and setting. We observe a time series X_t^{obs}, $t = 1, \ldots, n$, taking values in $\mathbb{R}^d$ for some fixed dimension $d \in \mathbb{N}$. The observed time series is assumed to arise from a known generative model with an unknown p-dimensional parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$. Each value of $\theta \in \Theta$ determines a law of a stochastic process. For any value of θ, we can use the generative model to simulate $s \in \mathbb{N}$ realizations of a length n time series $(X_{1:n}^{(r)}(\theta))_{r\in[s]}$. When we do not need to refer to a particular realization, we drop the superscript.

NOTATION: For real numbers $x, y \in \mathbb{R}$, denote $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. For a vector $x \in \mathbb{R}^d$, denote the ℓ^p norm by $\|x\|_p$ and the Euclidean norm by $\|x\| = \|x\|_2$. For a random vector X with distribution depending on $\theta \in \Theta$, $\|X\|_{\mathcal{L}^q(\theta)} = (\mathbb{E}_\theta \|X\|^q)^{1/q}$, where $\mathbb{E}_\theta(\cdot)$ denotes expectation with respect to this θ-dependent distribution. For any natural number $j \in \mathbb{N}$, denote $[j] = \{1, 2, \ldots, j\}$. For a sequence of random vectors X_t, $t = 1, \ldots, n$, each taking values in $\mathbb{R}^d$, we denote a subsequence as $X_{t_1:t_2} = (X_{t_1}, \ldots, X_{t_2})^\top$, which takes values in $\mathbb{R}^{(t_2-t_1+1)\times d}$ for some $t_1, t_2 \in [n]$.

Random Fourier features. The central idea is that we should estimate the p-dimensional parameter θ_0 by matching the values of a small number of random features of the simulated and observed data. Consider $k = 2p + 1$ randomly drawn functions $\varphi_1, \ldots, \varphi_k$ from a distribution D over a suitably nice function class $\mathcal{F}$. The k functions should be sampled independently of one another and of the data. Each function $\varphi_i : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}$ takes as input a length $m + 1$ subsequence of the d-dimensional time series, where m is chosen based on the dynamics of the generative model. For instance, if the model is an order m^* Markov process, then set $m = m^*$.

Let $x_1, \ldots, x_{m+1}$ be vectors in $\mathbb{R}^d$, and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathbb{R}^{(m+1)\times d}$. Here we focus on random Fourier features, though our framework can be used with different random features. Specifically, we consider random Fourier features of the form

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

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 all k of these functions by

$$\varphi = (\varphi_1, \ldots, \varphi_k), \qquad \textbf{[2]}$$

so that $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^k$. Define the i-th random feature of the observed and r-th simulated time series at time $t = m + 1, \ldots, n$ as

$$f_{t,i}^{\text{obs}} = \varphi_i(X_{t-m:t}^{\text{obs}}), \quad f_{t,i}^{(r)}(\theta) = \varphi_i\left(X_{t-m:t}^{(r)}(\theta)\right).$$

Similarly, write all k random features at time t as

$$\begin{aligned} f_t^{\text{obs}} &= \left(f_{t,1}^{\text{obs}}, \ldots, f_{t,k}^{\text{obs}}\right), \\ f_t^{(r)}(\theta) &= \left(f_{t,1}^{(r)}(\theta), \ldots, f_{t,k}^{(r)}(\theta)\right). \end{aligned} \qquad \textbf{[3]}$$

Estimators.

Time-average estimator. For processes that are stationary, or at least asymptotically mean stationary, it suffices to consider the time-average of $k = 2p + 1$ random features. Denote the time-averages of the observed and simulated random features by

$$\begin{aligned} F^{\text{obs}} &= \frac{1}{n-m} \sum_{t=m+1}^{n} f_t^{\text{obs}}, \\ \bar{F}^{\text{sim}}(\theta) &= \frac{1}{n-m} \sum_{t=m+1}^{n} \bar{f}_t^{\text{sim}}(\theta), \end{aligned} \qquad \textbf{[4]}$$

where $\bar{f}_t^{\text{sim}}(\theta) = \frac{1}{s}\sum_{r=1}^{s} f_t^{(r)}(\theta)$. The time-average estimator is given by

$$\hat{\theta}^{\text{TA}} = \underset{\theta\in\Theta}{\operatorname{argmin}} \left\|\hat{Q}_n^{\text{TA}}(\theta)\right\|, \qquad \textbf{[5]}$$

where the sample discrepancy function is defined as

$$\hat{Q}_n^{\text{TA}}(\theta) = F^{\text{obs}} - \bar{F}^{\text{sim}}(\theta). \qquad \textbf{[6]}$$

In practice, an optimization procedure is used to find the minimizer of $\left\|\hat{Q}_n^{\text{TA}}(\cdot)\right\|$.

Rolling-window estimator. For many nonstationary processes, a different approach is required because the time-average of the (time-varying) expectations of the random features may not converge to a limiting mean. Even when it does converge, the limiting mean may not uniquely identify each θ. Therefore, we introduce the following *rolling-window* estimator.

For some window size $w \in \mathbb{N}$, denote the rolling-window averages of the observed and simulated random features at time t by

$$F_t^{\text{obs}} = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee(m+1)}^{t} f_j^{\text{obs}}, \qquad \bar{F}_t^{\text{sim}}(\theta) = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee(m+1)}^{t} \bar{f}_j^{\text{sim}}(\theta), \tag{7}$$

where $\bar{f}_t^{\text{sim}}(\theta) = \frac{1}{s}\sum_{r=1}^{s} f_t^{(r)}(\theta)$. The rolling-window estimator is given by

$$\hat{\theta}^{\text{RW}} = \underset{\theta\in\Theta}{\operatorname{argmin}} \left|\hat{Q}_n^{\text{RW}}(\theta)\right|, \tag{8}$$

where the sample discrepancy function is defined as

$$\begin{aligned} \hat{Q}_n^{\text{RW}}(\theta) &= \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left[\left\|\Delta_{t-L}\right\|^2 + K_t(\theta)\right], \\ K_t(\theta) &= 2\left(\Delta_{t-L}\right)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta)\right] - \left[\Delta_{t-L}\right]\right), \\ \Delta_{t-L} &= F_{t-L}^{\text{obs}} - \bar{F}_{t-L}^{\text{sim}}(\theta), \end{aligned} \tag{9}$$

where $\tau \in \mathbb{N}$ is an initial time-offset and $L \in \mathbb{N}$ is a lag. An optimization procedure is used to find the minimizer of $\left|\hat{Q}_n^{\text{RW}}(\cdot)\right|$. We select $w = w_n$, $\tau = \tau_n$, and $L = L_n$ as follows, so that they grow with the sample size n. The lag L_n is selected as $\lceil \frac{1}{10}\log(n)^2 \rceil$, the window size w_n is selected as the integer within $[n^{\frac{1}{2}}, n^{\frac{3}{4}}]$ that minimizes the sum of squared distances $\sum_{t=m+1+L}^{n} \left\|F_{t-L}^{\text{obs}} - f_t^{\text{obs}}\right\|^2$, and the offset is selected as $\tau_n = w_n$.

Extensions. Some extensions are possible. First, both estimators can be generalized by replacing the Euclidean norm with a weighted norm, using a weight matrix given by the inverse of a suitable estimator of the long-run covariance matrix. Second, the sample averages used in both estimators can be replaced with different mean estimators.

Summary of theory. We prove that both estimators are consistent under the regularity conditions stated in the Supplementary Information. The key idea used is the fractal Whitney embedding prevalence theorem; see Theorem 9.1 in (13). This result allows us to embed the p-dimensional parameter space Θ into $\mathbb{R}^{2p+1}$ with "generic" choices of random features. (The exact definition of genericity is given in the Supplementary Information.) Another important part of our analysis is the representation of stochastic processes as functions of past noise inputs. This representation dates back to the work of (14) and (15), although its utility for statistical inference was first articulated by (16). Later work extended this theory to the nonstationary setting (17–20). Our analysis builds on ideas from this literature and may be of independent interest.

Data, Materials, and Software Availability. The generated data that support the findings of this study will be made publicly available and are available to editors and referees during peer review. The code will be made publicly available in a GitHub repository and is available to editors and referees during peer review.

ACKNOWLEDGMENTS. This work was supported by NSF grant DMS-2310834. git

1. SN Wood, Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466, 1102–1104 (2010).
2. MA Beaumont, Approximate bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41, 379–406 (2010).
3. K Cranmer, J Brehmer, G Louppe, The frontier of simulation-based inference. Proc. Natl. Acad. Sci. 117, 30055–30062 (2020).
4. A Zammit-Mangion, M Sainsbury-Dale, R Huser, Neural methods for amortized inference. Annu. Rev. Stat. Its Appl. 12, 311–335 (2025).
5. A Rahimi, B Recht, Random features for large-scale kernel machines. Adv. Neural Inf. Process. Syst. 20 (2007).
6. T Sauer, JA Yorke, M Casdagli, Embedology. J. Stat. Phys. 65, 579–616 (1991).
7. O Kallenberg, Foundations of Modern Probability. (Springer-Verlag, New York), Second edition, (2002).
8. RM May, Simple mathematical models with very complicated dynamics. Nature 261, 459–467 (1976).
9. RL Devaney, A First Course in Chaotic Dynamical Systems: Theory and Experiment. (Addison-Wesley, Reading, Massachusetts), (1992).
10. WO Kermack, AG McKendrick, A contribution to the mathematical theory of epidemics. Proc. Royal Soc. Lond. Ser. A 115, 700–721 (1927).
11. AJ Lotka, Elements of Physical Biology. (Williams & Wilkins, Baltimore), (1925).
12. V Volterra, Fluctuations in the abundance of a species considered mathematically. Nature 118, 558–560 (1926).
13. J Yorke, W Ott, Prevalence. Bull. Am. Math. Soc. 42, 263–290 (2005).
14. M Rosenblatt, Independence and dependence. Proc. 4th Berkeley Symp. on Math. Stat. Prob. 2, 431–443 (1961).
15. N Wiener, Nonlinear Problems in Random Theory. (MIT Press, Cambridge, MA), (1966).
16. WB Wu, Nonlinear system theory: another look at dependence. Proc. Natl. Acad. Sci. 102, 14150–14154 (2005).
17. Z Zhou, WB Wu, Simultaneous inference of linear models with time varying coefficients. J. Royal Stat. Soc. Ser. B: Stat. Methodol. 72, 513–531 (2010).
18. R Dahlhaus, S Richter, WB Wu, Towards a general theory for nonlinear locally stationary processes. Bernoulli 25, 1013–1044 (2019).
19. F Mies, A Steland, Sequential gaussian approximation for nonstationary time series in high dimensions. Bernoulli 29, 3114–3140 (2023).
20. F Mies, Functional estimation and change detection for nonstationary time series. J. Am. Stat. Assoc. 118, 1011–1022 (2023).

Supporting Information Appendix (SI)

Contents

1. Theory

In this section, we present the assumptions and theoretical results for the estimators.

1.1. Noise inputs.

Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be an iid sequence of random variables. Let $\left(\varepsilon_i^{(r)}\right)_{i\in\mathbb{Z}}$, $r = 0, 1, \ldots, s$, be iid copies of $(\varepsilon_i)_{i\in\mathbb{Z}}$. Define the sequence of noise inputs up to time t as

$$\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots), \qquad \textbf{[10]}$$

and define $\boldsymbol{\varepsilon}_t^{(r)}$, $r = 0, 1, \ldots, s$, analogously. For each time $t = 1, \ldots, n$, we express X_t^{obs} and each $X_t^{(r)}(\theta)$ as functions of one of these noise inputs. In practice, simulations only use finitely many noise inputs. However, considering a countably infinite sequence of noise inputs is convenient for several reasons. For instance, this allows us to accommodate sequences of simulation models which are approximations to the true data generating process, where the approximation improves by using a growing number of noise inputs in the burn-in period.

Let $(\varepsilon_i^*)_{i\in\mathbb{Z}}$ be an iid copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$. For $t \in \mathbb{Z}$, $j \in \mathbb{N}_0$, define

$$\tilde{\boldsymbol{\varepsilon}}_{t,j} = \left(\varepsilon_t, \ldots, \varepsilon_{t-j+1}, \varepsilon_{t-j}^*, \varepsilon_{t-j-1}, \ldots\right), \qquad \textbf{[11]}$$

and define $\tilde{\boldsymbol{\varepsilon}}_{t,j}^{(r)}$, $r = 0, 1, \ldots, s$, analogously. $\tilde{\boldsymbol{\varepsilon}}_{t,j}$ is the sequence of noise inputs $\boldsymbol{\varepsilon}_t$ with the innovation from j steps in the past, ε_{t-j}, replaced by an iid copy ε_{t-j}^*. Intuitively, the output of the generative model at time t changes if we replace the noise input from j steps ago with an iid copy because the process at time t is a function of the past noise inputs. Our central assumption is that the generative model's current output becomes increasingly insensitive to the j-th noise input in the past as j grows.

To formalize this, we use the physical dependence measure of (1). Under certain regularity conditions, the physical dependence measure conditions we impose imply and are implied by β-mixing and strong mixing conditions (2, 3).

1.2. Results for time-average estimator.

We state sufficient conditions for $\hat{\theta}^{\text{TA}}$ from Eq. (5) to be consistent. The main result is stated in Theorem 3. In a separate paper focused on statistical inference, we establish asymptotic normality under slightly stronger conditions, which implies a convergence rate of $O(1/\sqrt{n})$.

In this setting, we use long-run asymptotics. That is, n increasing corresponds to observing the process further into the future. We allow the observed and simulated time series to be generated by approximations to a limiting data generating process. This accommodates many settings with imperfect observations (e.g., observing residuals rather than errors), discretization errors, unknown initial values, and growing burn-in periods.

For each $t \in \mathbb{N}$, $i \in \mathbb{N}$, $r = 0, 1, \ldots, s$, the measurable mapping $G_t^{(i,r)}$ is from $\mathbb{R}^\infty \times \Theta$ to $\mathbb{R}^d$ and $(G_t^{(i,r)}(\boldsymbol{\varepsilon}_j, \theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process. The observed and simulated time series may be nonstationary because the mappings may change over time $t \in \mathbb{N}$. Throughout, we endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections. The mapping $G_t^{(i,r)}$ can be understood as an approximation to some limiting mapping $G_t^{(r)}$, where the approximation improves as i grows and the dependence on r allows for random initial conditions. See Assumption 2 and the following discussion for more details. The superscript may be ignored if approximations are unnecessary.

The index $n \in \mathbb{N}$ is linked to both the approximation and the sample size.

Assumption 1. *The observed and simulated time series of length $n \in \mathbb{N}$ are*

$$X_t^{\text{obs}} = G_t^{(n,0)}(\varepsilon_t^{(0)}, \theta_0), \quad X_t^{(r)}(\theta) = G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta), \quad t = 1, \ldots, n, \qquad \textbf{[12]}$$

for all $r = 1, \ldots, s$ and $\theta \in \Theta$.

Oftentimes, we can take the observed time series to be generated by the limiting mapping, i.e. $G_t^{(n,0)} \equiv G_t^{(0)}$ for all $n \in \mathbb{N}$. The following assumption imposes the convergence to limiting mappings as the approximation level grows. Note that no rate of convergence is required. We write ε_0 because it has the same distribution as each ε_t.

Assumption 2. *For each $t \in \mathbb{N}$ and $r = 0, 1, \ldots, s$, there exists a measurable mapping $G_t^{(r)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$ such that $(G_t^{(r)}(\varepsilon_j, \theta))_{j \in \mathbb{Z}}$ is a stationary ergodic process, and for some $q > 2$, as $n \to \infty$, we have*

$$\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{t=1}^{n} \left\| G_t^{(n,r)}(\varepsilon_0, \theta) - G_t^{(r)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} \right) \to 0.$$

Under Assumption 1, the $k = 2p + 1$ time-averages from Eq. (4) can be written as

$$\begin{aligned} F^{\text{obs}} &= \frac{1}{n-m} \sum_{t=m+1}^{n} \varphi \left(\left[G_{t-m}^{(n,0)}(\varepsilon_{t-m}^{(0)}, \theta_0), \ldots, G_t^{(n,0)}(\varepsilon_t^{(0)}, \theta_0) \right]^\top \right), \\ \bar{F}^{\text{sim}}(\theta) &= \frac{1}{n-m} \sum_{t=m+1}^{n} \left[\frac{1}{s} \sum_{r=1}^{s} \varphi \left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right]. \end{aligned} \qquad \textbf{[13]}$$

Denote by $P_{\theta,t}^{(r)}$ the distribution of $\left[G_{t-m}^{(r)}(\varepsilon_{-m}, \theta), \ldots, G_t^{(r)}(\varepsilon_0, \theta) \right]^\top$, where m is the number of lags in the random features φ from Eq. (2). We impose asymptotic mean stationarity by requiring the time-average of the distributions $(P_{\theta,t}^{(r)})_{t \in \mathbb{N}}$ to converge weakly to a limiting distribution P_θ which depends neither on t nor r, *uniformly* over $\theta \in \Theta$. For example, this allows us to consider discrete-time dynamical systems with unknown initial value.

Assumption 3. *For all $\theta \in \Theta$, there exists a limiting distribution P_θ, such that, for all $r = 0, 1, \ldots, s$, we have, as $n \to \infty$,*

$$\sup_{\theta \in \Theta} \left| \int_{\mathbb{R}^{(m+1) \times d}} \phi(x) \, d\bar{P}_{\theta,n}^{(r)}(x) - \int_{\mathbb{R}^{(m+1) \times d}} \phi(x) \, dP_\theta(x) \right| \to 0,$$

for all continuous, bounded functions $\phi : \mathbb{R}^{(m+1) \times d} \to \mathbb{R}$, where $\bar{P}_{\theta,n}^{(r)} = \frac{1}{n-m} \sum_{t=m+1}^{n} P_{\theta,t}^{(r)}$.

Next, we introduce mappings $\Phi : \mathbb{R}^p \to \mathbb{R}^k$, whose restrictions to Θ satisfy

$$\Phi|_\Theta(\theta) = \int_{\mathbb{R}^{(m+1) \times d}} \varphi(x) \, dP_\theta(x), \quad \theta \in \Theta. \qquad \textbf{[14]}$$

Under Assumptions 1-3, the limiting expectation values of the $k = 2p + 1$ random features of the time series coincide with Φ restricted to Θ. Roughly speaking, we would like to say that, on Θ, for "generic" choices of the $k = 2p + 1$ random features, we have that Φ is C^1, Φ^{-1} exists, and Φ^{-1} is also C^1. For more details, see the proof of Theorem 2.

(4) introduce the following *genericity axioms*, which serve as desiderata for a notion of "generic". For a topological vector space V, we desire that:

1. A generic subset of V is dense in V.
2. If $F \supset E$ and E is generic, then F is generic.
3. A countable intersection of generic sets is generic.
4. Every translate of a generic set is generic.
5. A subset E of finite-dimensional Euclidean space is generic if and only if E is a set of full Lebesgue measure.

The topological notion of genericity is based on the idea of *residual sets*, defined as a countable intersection of open, dense sets. The Baire category theorem states that residual subsets of Baire spaces are dense. However, the Lebesgue measure of these residual sets can be zero; see (4) for examples. As a result, the topological notion of genericity does not satisfy Axiom 5. Hence, it would be of interest to use a measure-theoretic notion of "generic".

As discussed in (4–6), the difficulty with statements asserting that a property holds for "almost every" smooth map is that the space of all smooth maps is an infinite-dimensional space. The idea of probability one in finite dimensions does not extend in any canonical way to infinite-dimensional spaces. Nevertheless, it is desirable to formalize the idea that a property holds for "almost all" maps. To do this, we adopt the notion of *prevalence* from (4), which therein is shown to satisfy all of the genericity axioms.

Definition 1. (Prevalence (4)). Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be prevalent if there exists some Borel measure μ on V such that:

1. $0 < \mu(C) < \infty$ for some compact subset C of V.
2. The set $E + v$ has full μ-measure (that is, the complement of $E + v$ has measure zero) for all $v \in V$.

In this literature, it is common to say that, if $E \subset V$ contains a prevalent Borel set, then "almost every" element of V lies in E. In this sense, "almost every" smooth map from $\mathbb{R}^p$ to $\mathbb{R}^k$ has the following properties.

Theorem 1. (Fractal Whitney Embedding Prevalence Theorem (5)). *Let Θ be a compact subset of $\mathbb{R}^p$ of box-counting dimension p', and let k be an integer greater than $2p'$. For almost every C^1 smooth map $\Phi : \mathbb{R}^p \to \mathbb{R}^k$:*

1. *Φ is one-to-one on Θ.*
2. *Φ is an immersion on each compact subset of a smooth manifold contained in Θ.*

We impose the following conditions so that we may apply Theorem 1. Let $\mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$ denote the space of probability measures on $\mathbb{R}^{(m+1)\times d}$.

Assumption 4. *The following conditions hold:*

1. *Θ is a compact subset of $\mathbb{R}^p$ with non-empty interior.*
2. *The map $\theta \mapsto P_\theta$ is a bijection.*
3. *For some open set $O \supset \Theta$, the map $\theta \mapsto P_\theta$ admits an extension $\mathbf{p} : O \to \mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$ that is Fréchet C^1 on O with respect to the total variation norm.*

That is, there exists an open set $O \subset \mathbb{R}^p$ with $\Theta \subset O$ and a map $\mathbf{p} : O \to \mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$, such that $\mathbf{p}(\theta) = P_\theta$ for all $\theta \in \Theta$, and $\mathbf{p}$ is Fréchet C^1 on O with respect to the total variation norm. We impose structure on the extension $\mathbf{p}$ from Assumption 4 by linking it to Φ through the definition of the function class

$$\mathcal{C}_\varphi = \left\{ \Phi : \mathbb{R}^p \to \mathbb{R}^k \;\middle|\; \Phi|_O(\theta) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\mathbf{p}(\theta)(x),\ \Phi \text{ is } C^1 \text{ on } \mathbb{R}^p \setminus \Theta \right\},$$

and let $\mathcal{C} = \bigcup_{\varphi \in \mathcal{F}^k} \mathcal{C}_\varphi$. In words, $\mathcal{C}$ consists of all functions $\Phi : \mathbb{R}^p \to \mathbb{R}^k$ whose restriction to the open set $O \subset \mathbb{R}^p$ from Assumption 4 coincides with the expectations of the k random features with respect to $\mathbf{p}(\theta), \theta \in O$, for some $\varphi \in \mathcal{F}^k$, and which admit a C^1 extension to the rest of $\mathbb{R}^p$, the existence of which is ensured by the (7) extension theorem and its generalizations.

Theorem 2. *Suppose Assumptions 1-4 hold. Then the limiting expectation values of the k random features are C^1-smooth in the parameter values, so $\mathcal{C} \subset C^1(\mathbb{R}^p, \mathbb{R}^k)$. Furthermore, once $k \geq 2p+1$, almost every C^1 smooth map from $\mathbb{R}^p$ to $\mathbb{R}^k$ is one-to-one on Θ, and is an immersion on each compact subset of a smooth manifold contained in Θ.*

To simplify the presentation of the following results, we make the assumption that the $k = 2p+1$ random features yield such a function.

Assumption 5. *The function Φ from* Eq. (14) *is one-to-one on Θ, and is an immersion on each compact subset of a smooth manifold contained in Θ.*

Therefore, if we knew Φ, then we could smoothly translate between the values of θ and expectations of the random features $\Phi(\theta)$.

The following stochastic equicontinuity-type assumption is useful for establishing the consistency of the estimator.

Assumption 6. *For all $i \in \mathbb{N}$, all $r = 0, 1, \ldots, s$, and all $\epsilon > 0$, there exists an $\eta_G > 0$ such that*

$$\sup_{t\in\mathbb{N}} \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\leq\epsilon}} \left\| G_t^{(i,r)}(\varepsilon_0, \theta) - G_t^{(i,r)}(\varepsilon_0, \theta') \right\| \right) < \eta_G,$$

with $\eta_G \to 0$ as $\epsilon \to 0$, where the expectation $\mathbb{E}(\cdot)$ is taken with respect to the law of the noise inputs.

The next assumption controls the temporal dependence uniformly over Θ, so that we can estimate the expectations of the random features based on their time-averages. Specifically, we impose a uniform moment bound and a uniform polynomial decay of the physical dependence measure of (1).

Assumption 7. *There exist $\Psi > 0$, $\beta > 2$ such that, for some $q > 2$, all $j \in \mathbb{N}$, all $i \in \mathbb{N}$, and all $r = 0, 1, \ldots, s$, we have*

$$\sup_{\theta\in\Theta}\sup_{t\in\mathbb{N}} \left\| G_t^{(i,r)}(\varepsilon_0, \theta) - G_t^{(i,r)}(\tilde{\varepsilon}_{0,j}, \theta) \right\|_{\mathcal{L}^q(\theta)} \le \Psi(j \vee 1)^{-\beta},$$

$$\sup_{\theta\in\Theta}\sup_{t\in\mathbb{N}} \left\| G_t^{(i,r)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} \le \Psi,$$

where $\tilde{\varepsilon}_{0,j}$ is from Eq. (11).

We now establish the consistency of the time-average estimator $\hat{\theta}^{\mathrm{TA}}$ from Eq. (5).

Theorem 3. *If Assumptions 1-7 hold, then $\hat{\theta}^{\mathrm{TA}} \xrightarrow{p} \theta_0$.*

1.3. Results for rolling-window estimator. We present assumptions that guarantee the consistency of the rolling-window estimator $\hat{\theta}^{\mathrm{RW}}$ from Eq. (8). The main result is Theorem 6. In a separate manuscript focused on statistical inference, we prove asymptotic normality under slightly stronger conditions, which implies a $O(1/\sqrt{n})$ rate of convergence.

In this section, we rescale time to the unit interval and use infill asymptotics as in the literature on locally stationary processes by (8, 9). As n grows, an increasing number of observations related to each local structure of the process become available. The rolling-window estimator introduced here can be used to fit models with very general forms of nonstationarity, such as ordinary differential equations (ODEs), state-space models with covariates and control inputs, and structural time series models with complicated trends, seasonality, change-points, and "rough" nonstationarity.

The observed and simulated time series may be generated by approximations to a limiting data generating process. For example, this distinction allows the observed time series to be based on observed residuals rather than errors, or to consist of imperfect observations of an exact solution of an ODE. Similarly, the simulated time series can be based on an ODE solver with observational noise processes that have growing burn-in periods.

We represent the generative model using measurable mappings. For each $u \in [0,1]$, $i \in \mathbb{N}$, $r = 0, 1, \ldots, s$, the mapping $G_u^{(i,r)}$ is from $\mathbb{R}^\infty \times \Theta$ to $\mathbb{R}^d$ and $(G_u^{(i,r)}(\varepsilon_j, \theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process. The time series may be nonstationary because the mappings may change over rescaled time $u \in [0,1]$. The mapping $G_u^{(i,r)}$ is an approximation to a limiting mapping. When an approximation is not needed, the superscript can be ignored.

The index $n \in \mathbb{N}$ is linked to the sample size and approximation.

Assumption 8. *The observed and simulated time series of length $n \in \mathbb{N}$ are*

$$X_t^{\mathrm{obs}} = G_{t/n}^{(n,0)}(\varepsilon_t^{(0)}, \theta_0), \quad X_t^{(r)}(\theta) = G_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta), \quad t = 1, \ldots, n, \qquad [15]$$

for all $r = 1, \ldots, s$ and $\theta \in \Theta$.

In many situations, we can take the observed time series to be generated by the limiting mapping, i.e. $G_u^{(n,0)} \equiv G_u^{(0)}$ for all $u \in [0,1]$. The next assumption imposes convergence to the limiting mapping, with no required convergence rate. Note that ε_0 has the same distribution as ε_t for all $t \in \mathbb{Z}$.

Assumption 9. *For each $u \in [0,1]$, there exists a measurable mapping $G_u : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$ such that $(G_u(\varepsilon_j, \theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process, and for some $q > 2$ and all $r = 0, 1, \ldots, s$, as $i \to \infty$, we have*

$$\sup_{\theta\in\Theta}\sup_{u\in[0,1]} \left\| G_u^{(i,r)}(\varepsilon_0, \theta) - G_u(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} \to 0.$$

Under Assumption 8, the observed and simulated rolling-window random features at time t can be written as

$$F_t^{\mathrm{obs}} = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee(m+1)}^{t} \varphi\left(\left[G_{(j-m)/n}^{(n,0)}(\varepsilon_{j-m}^{(0)}, \theta_0), \ldots, G_{j/n}^{(n,0)}(\varepsilon_j^{(0)}, \theta_0) \right]^\top \right),$$

$$\bar{F}_t^{\mathrm{sim}}(\theta) = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee(m+1)}^{t} \left[\frac{1}{s} \sum_{r=1}^{s} \varphi\left(\left[G_{(j-m)/n}^{(n,r)}(\varepsilon_{j-m}^{(r)}, \theta), \ldots, G_{j/n}^{(n,r)}(\varepsilon_j^{(r)}, \theta) \right]^\top \right) \right].$$

For all $u \in [0,1]$, denote by $P_{\theta,u}$ the distribution of $[G_u(\varepsilon_{-m}, \theta), \ldots, G_u(\varepsilon_0, \theta)]^\top$, where m is the number of lags used in the random features. We introduce mappings $\Phi_u : \mathbb{R}^p \to \mathbb{R}^k$, $u \in [0,1]$, whose restrictions to Θ satisfy

$$\Phi_u|_\Theta(\theta) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, dP_{\theta,u}(x). \qquad [16]$$

Under Assumptions 8 and 9, the expectation values of the $k = 2p+1$ random features at rescaled time $u \in [0,1]$ coincide with Φ_u restricted to Θ.

Unfortunately, in general, we cannot embed Θ into $\mathbb{R}^k$ using Φ_u at one particular rescaled time $u \in [0,1]$. To see this, consider a family of stochastic processes defined by $X_t(\theta) = \cos(\theta 2\pi t/n) + \varepsilon_t$, $t = 1, \ldots, n$, where each $\varepsilon_t \overset{\text{iid}}{\sim} N(0,1)$ and $\theta \in [0,10]$. For $u \in [0,1]$ and $m = 0$, denote by $P_{\theta,u}$ the distribution of $\cos(\theta 2\pi u) + \varepsilon_t$, i.e. $P_{\theta,u} = N(\cos(\theta 2\pi u), 1)$. Clearly, for two different parameters $\theta_1 \neq \theta_2$, it is possible for $P_{\theta_1,u} = P_{\theta_2,u}$. Of course, this also occurs in more complicated dynamic models, such as differential equations. To deal with these situations, we make the following reparametrization assumption.

Assumption 10. *For all $u \in [0,1]$, there exists a Lipschitz mapping $g_u : \Theta \to \tilde{\Theta}_u$ defined by $g_u(\theta) = \tilde{\theta}_u$ where $\tilde{\Theta}_u = \{g_u(\theta) : \theta \in \Theta\}$ is a subset of $\mathbb{R}^{\tilde{p}}$ for some $\tilde{p} \in \mathbb{N}$, and there exists a function $\tilde{G}_u : \mathbb{R}^\infty \times \tilde{\Theta}_u \to \mathbb{R}^d$ such that*

$$\tilde{G}_u(\varepsilon_0, \tilde{\theta}_u) = G_u(\varepsilon_0, \theta).$$

For each fixed $u \in [0,1]$, $(\tilde{G}_u(\varepsilon_j, \tilde{\theta}_u))_{j \in \mathbb{Z}}$ is a stationary ergodic process.

In the previous example, we may take $g_u(\theta) = \cos(\theta 2\pi u)$. Similar ideas apply for our experiments with the rolling-window estimator (e.g., with ODEs). When $\theta \mapsto P_{\theta,u}$ is a bijection for all $u \in [0,1]$, we may always take $g_u(\theta) = \theta$.

Going forward, let $\tilde{P}_{\tilde{\theta}_u,u}$ denote the distribution of $\left[\tilde{G}_u(\varepsilon_{-m}, \tilde{\theta}_u), \ldots, \tilde{G}_u(\varepsilon_0, \tilde{\theta}_u)\right]^\top$. Now, we introduce the mappings $\tilde{\Phi}_u : \mathbb{R}^{\tilde{p}} \to \mathbb{R}^k$ whose restrictions to $\tilde{\Theta}_u$ satisfy

$$\tilde{\Phi}_u|_{\tilde{\Theta}_u}(\tilde{\theta}_u) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\tilde{P}_{\tilde{\theta}_u,u}(x). \qquad \textbf{[17]}$$

Under Assumptions 8-10, the limiting local expectation values at rescaled time $u \in [0,1]$ of the $k = 2p+1$ random features of the time series are given by $\tilde{\Phi}_u(\tilde{\theta}_u)$, $\tilde{\theta}_u \in \tilde{\Theta}_u$. We would like to say that, on $\tilde{\Theta}_u$, for "generic" choices of the $k = 2p+1$ random features, we have that $\tilde{\Phi}_u$ is C^1, its inverse exists, and is also C^1. Again, we adopt the notion of prevalence from Definition 1 and use the fractal Whitney embedding prevalence theorem. For convenience, we restate the theorem here in the notation of this section.

Theorem 4. (Fractal Whitney Embedding Prevalence Theorem (5)). *Let $\tilde{\Theta}_u$ be a compact subset of $\mathbb{R}^{\tilde{p}}$ of box-counting dimension p, and let k be an integer greater than $2p$. For almost every C^1 smooth map $\tilde{\Phi}_u : \mathbb{R}^{\tilde{p}} \to \mathbb{R}^k$:*

1. *$\tilde{\Phi}_u$ is one-to-one on $\tilde{\Theta}_u$.*
2. *$\tilde{\Phi}_u$ is an immersion on each compact subset of a smooth manifold contained in $\tilde{\Theta}_u$.*

We impose the following conditions so that we may apply Theorem 4 at each rescaled time $u \in [0,1]$. Denote by $\mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$ the space of probability measures on $\mathbb{R}^{(m+1)\times d}$.

Assumption 11. *The following conditions hold for all rescaled times $u \in [0,1]$:*

1. *$\tilde{\Theta}_u$ is a compact subset of $\mathbb{R}^{\tilde{p}}$, and Θ is a compact subset of $\mathbb{R}^p$ with non-empty interior.*
2. *The map $\tilde{\theta}_u \mapsto \tilde{P}_{\tilde{\theta}_u,u}$ is a bijection.*
3. *For some open set $\tilde{O}_u \supset \tilde{\Theta}_u$, the map $\tilde{\theta}_u \mapsto \tilde{P}_{\tilde{\theta}_u,u}$ admits an extension $\tilde{\mathbf{p}}_u : \tilde{O}_u \to \mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$ that is Fréchet C^1 on $\tilde{O}_u$ with respect to the total variation norm.*

That is, for all $u \in [0,1]$, there exists an open set $\tilde{O}_u \subset \mathbb{R}^{\tilde{p}}$ with $\tilde{\Theta}_u \subset \tilde{O}_u$ and a map $\tilde{\mathbf{p}}_u : \tilde{O}_u \to \mathcal{P}\left(\mathbb{R}^{(m+1)\times d}\right)$, such that $\tilde{\mathbf{p}}_u(\tilde{\theta}_u) = \tilde{P}_{\tilde{\theta}_u,u}$ for all $\theta \in \Theta$, and $\tilde{\mathbf{p}}_u$ is Fréchet C^1 on O with respect to the total variation norm. We impose structure on these extensions $\tilde{\mathbf{p}}_u$, $u \in [0,1]$, by linking them to $\tilde{\Phi}_u$, $u \in [0,1]$, through the definition of the function classes

$$\tilde{\mathcal{C}}_{u,\varphi} = \left\{ \tilde{\Phi}_u : \mathbb{R}^{\tilde{p}} \to \mathbb{R}^k \;\middle|\; \tilde{\Phi}_u|_{\tilde{O}_u}(\tilde{\theta}_u) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x) d\tilde{\mathbf{p}}_u(\tilde{\theta}_u)(x), \tilde{\Phi}_u \text{ is } C^1 \text{ on } \mathbb{R}^{\tilde{p}} \setminus \tilde{\Theta}_u \right\},$$

for $u \in [0,1]$, and let $\tilde{\mathcal{C}}_u = \bigcup_{\varphi \in \mathcal{F}^k} \tilde{\mathcal{C}}_{u,\varphi}$. In words, $\tilde{\mathcal{C}}_u$ is the collection of all functions $\tilde{\Phi}_u : \mathbb{R}^{\tilde{p}} \to \mathbb{R}^k$ whose restriction to the open set $\tilde{O}_u \subset \mathbb{R}^{\tilde{p}}$ from Assumption 11 coincides with the local expectations of the k random features with respect to $\tilde{\mathbf{p}}_u(\tilde{\theta}_u)$, $\tilde{\theta}_u \in \tilde{O}_u$, for some $\varphi \in \mathcal{F}^k$, and which admit a C^1 extension to the rest of $\mathbb{R}^{\tilde{p}}$, the existence of which is ensured by the (7) extension theorem and its generalizations.

Theorem 5. *Suppose Assumptions 8-11 hold. Then, for all $u \in [0,1]$, the limiting local expectation values at rescaled time u of the k random features are C^1-smooth in the parameter values, so $\tilde{\mathcal{C}}_u \subset C^1(\mathbb{R}^{\tilde{p}}, \mathbb{R}^k)$. Furthermore, for all $u \in [0,1]$, once $k \geq 2p+1$, almost every C^1 smooth map from $\mathbb{R}^{\tilde{p}}$ to $\mathbb{R}^k$ is one-to-one on $\tilde{\Theta}_u$, and is an immersion on each compact subset of a smooth manifold contained in $\tilde{\Theta}_u$.*

To simplify the presentation of the results below, we make the assumption that the $k = 2p+1$ random features yield such a function.

Assumption 12. *For each $u \in [0,1]$, the mapping $\tilde{\Phi}_u$ from* Eq. (17) *is one-to-one on $\tilde{\Theta}_u$, and is an immersion on each compact subset of a smooth manifold contained in $\tilde{\Theta}_u$.*

Next, we assume that the paths of distributions $(P_{\theta_1,u})_{u\in[0,1]}$, $(P_{\theta_2,u})_{u\in[0,1]}$ are sufficiently distinguishable, in some sense, for any two different parameters $\theta_1, \theta_2 \in \Theta$. This is a mild condition, as we still allow $P_{\theta_1,u} = P_{\theta_2,u}$ for all but a measurable amount of rescaled times.

Assumption 13. *For all $\theta_1, \theta_2 \in \Theta$, if $\theta_1 \neq \theta_2$, then there exists a subset $\mathcal{U}_{\theta_1,\theta_2} \subseteq [0,1]$ with positive Lebesgue measure such that $P_{\theta_1,u} \neq P_{\theta_2,u}$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$.*

We have stated the core assumptions, so we may now sketch the main ideas of the proof of Theorem 6, which shows the consistency of $\hat{\theta}^{\mathrm{RW}}$. By Assumption 13, for any two different parameters $\theta_1, \theta_2 \in \Theta$, we have that $P_{\theta_1,u} \neq P_{\theta_2,u}$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$. Since $P_{\theta_1,u} \neq P_{\theta_2,u}$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$, we also have that $\tilde{P}_{g_u(\theta_1),u} \neq \tilde{P}_{g_u(\theta_2),u}$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$ because $\tilde{P}_{g_u(\theta),u} = P_{\theta,u}$ is the distribution of

$$[\tilde{G}_u(\varepsilon_{-m}, g_u(\theta)), \ldots, \tilde{G}_u(\varepsilon_0, g_u(\theta))]^\top = [G_u(\varepsilon_{-m}, \theta), \ldots, G_u(\varepsilon_0, \theta)]^\top.$$

Therefore, for all $u \in \mathcal{U}_{\theta_1,\theta_2}$, we have that $g_u(\theta_1) \neq g_u(\theta_2)$ because $\tilde{\theta}_u \mapsto \tilde{P}_{\tilde{\theta}_u,u}$ is a bijection for all $u \in [0,1]$ by Assumption 11. We also have that $\tilde{\Phi}_u(g_u(\theta_1)) \neq \tilde{\Phi}_u(g_u(\theta_2))$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$, because by Assumption 12, $\tilde{\theta}_u \mapsto \tilde{\Phi}_u(\tilde{\theta}_u)$ is one-to-one for all $u \in [0,1]$. This also implies that $\Phi_u(\theta_1) \neq \Phi_u(\theta_2)$ for all $u \in \mathcal{U}_{\theta_1,\theta_2}$ by the definitions of Φ_u and $\tilde{\Phi}_u$ from Eq. (16) and Eq. (17), respectively, because $\tilde{P}_{g_u(\theta),u} = P_{\theta,u}$ as discussed above. Thus, if Φ_u were known, we could distinguish θ_1 and θ_2 by looking at the expectation values of the random features $\Phi_u(\theta_1)$ and $\Phi_u(\theta_2)$ at rescaled times $u \in \mathcal{U}_{\theta_1,\theta_2}$.

We use the following stochastic equicontinuity-type assumption to establish the consistency of the estimator.

Assumption 14. *For all $i \in \mathbb{N}$, all $r = 0, 1, \ldots, s$, and all $\epsilon > 0$, there exists an $\eta_G > 0$ such that*

$$\sup_{u\in[0,1]} \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\leq\epsilon}} \left\| G_u^{(i,r)}(\varepsilon_0,\theta) - G_u^{(i,r)}(\varepsilon_0,\theta') \right\|^2 \right) < \eta_G,$$

with $\eta_G \to 0$ as $\epsilon \to 0$, where the expectation $\mathbb{E}(\cdot)$ is taken with respect to the law of the noise inputs.

The next assumptions control the temporal dependence and nonstationarity uniformly over Θ, so that we can estimate the expectations of the random features $\Phi_u(\theta)$ at all $u \in [0,1]$ and $\theta \in \Theta$. First, we impose a uniform moment bound and a uniform exponential decay of the physical dependence measure of (1).

Assumption 15. *There exist constants $\rho \in (0,1)$, $\Lambda > 0$, such that, for some $q > 2$, all $j \in \mathbb{N}$, all $i \in \mathbb{N}$, and all $r = 0, 1, \ldots, s$, we have*

$$\sup_{\theta\in\Theta} \sup_{u\in[0,1]} \left\| G_u^{(i,r)}(\varepsilon_0,\theta) - G_u^{(i,r)}(\tilde{\varepsilon}_{0,j},\theta) \right\|_{\mathcal{L}^q(\theta)} \leq \Lambda\rho^j,$$

$$\sup_{\theta\in\Theta} \sup_{u\in[0,1]} \left\| G_u^{(i,r)}(\varepsilon_0,\theta) \right\|_{\mathcal{L}^q(\theta)} \leq \Lambda,$$

where $\tilde{\varepsilon}_{0,j}$ is from Eq. (11).

Second, we impose the following regularity condition to control the nonstationarity, which allows for smooth changes, abrupt change-points, and "rough" nonstationarity. Define

$$\left\| \left(G_u^{(i,r)}(\varepsilon_0,\theta) \right)_u \right\|_{\kappa\text{-var},\mathcal{L}^q(\theta)} = \sup_{\substack{0=u_0<\ldots<u_\ell=1 \\ \ell\in\mathbb{N}}} \left(\sum_{j=1}^{\ell} \left\| G_{u_j}^{(i,r)}(\varepsilon_0,\theta) - G_{u_{j-1}}^{(i,r)}(\varepsilon_0,\theta) \right\|_{\mathcal{L}^q(\theta)}^{\kappa} \right)^{\frac{1}{\kappa}}.$$

Assumption 16. *For some $q > 2$, some $\kappa \in [1,4)$, all $i \in \mathbb{N}$, and all $r = 0, 1, \ldots, s$, the constant $\Lambda > 0$ from Assumption 15 also satisfies*

$$\sup_{\theta\in\Theta} \sup_{u\in[0,1]} \left\| G_u^{(i,r)}(\varepsilon_0,\theta) \right\|_{\mathcal{L}^q(\theta)} + \sup_{\theta\in\Theta} \left\| \left(G_u^{(i,r)}(\varepsilon_0,\theta) \right)_u \right\|_{\kappa\text{-var},\mathcal{L}^q(\theta)} \leq \Lambda.$$

Next, we specify the rates of growth for the window size w_n, the offset τ_n, and the lag L_n from Eq. (9). This ensures that we can improve our estimates of the local expectation values of the random features at $u \in [0,1]$ as we gain more observations near rescaled time u.

Assumption 17. *Recall $q > 2$ from Assumption 15 and $\kappa \in [1,4)$ from Assumption 16. Assume that q, κ, $(w_n)_{n\in\mathbb{N}}$, $(\tau_n)_{n\in\mathbb{N}}$, and $(L_n)_{n\in\mathbb{N}}$ satisfy:*

1. $\tau_n = o(n)$.

2. $w_n \gg n^{\frac{2}{q}}$, $w_n = o(n)$, and $w_n n^{\max(1-2/\kappa,0)} = o(n)$.

3. $L_n \gg \log(n)^{1+a}$ for some $a \in (0,1)$, $L_n \ll \tau_n$, $L_n \ll n^{\frac{1}{\kappa}-\frac{1}{4}}$, and $L_n = o(n^{\frac{1}{2\kappa}})$.

We now establish the consistency of the rolling-window estimator $\hat{\theta}^{\mathrm{RW}}$ from Eq. (8).

Theorem 6. *If Assumptions 8-17 hold, then* $\hat{\theta}^{\mathrm{RW}} \xrightarrow{p} \theta_0$.

2. Proofs

2.1. Proof of Theorem 1.

See the proof of Theorem 2.3 in (5).

□

2.2. Proof of Theorem 2.

Step 1: Limiting expectations. We claim that under Assumptions 1-3, $\forall \theta \in \Theta$, the limiting expectation values of the k random features of the time series are given by

$$\Phi(\theta) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, dP_\theta(x).$$

That is, for all $\theta \in \Theta$ and all $r \in \{0, 1, \ldots, s\}$, we have, as $n \to \infty$,

$$\left\| \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi \left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] - \Phi(\theta) \right\| = o(1). \qquad \textbf{[18]}$$

Note that the time series have this causal representation by Assumption 1.

Observe that

$$\begin{aligned}
&\left\| \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi \left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] - \Phi(\theta) \right\| \\
&\overset{(1)}{\leq} \left\| \begin{array}{l} \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi \left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] \\ - \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi \left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] \end{array} \right\| \\
&+ \left\| \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi \left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] - \Phi(\theta) \right\| \\
&\overset{(2)}{=} o(1),
\end{aligned}$$

where (1) is from adding and subtracting the corresponding term with the limiting mapping and by applying the triangle inequality, and (2) holds because both terms are $o(1)$. Specifically, the second term is $o(1)$ by upper bounding the ℓ_2 norm $\|\cdot\|$ by the ℓ_1 norm (i.e. over the k random features) using the monotonicity of norms, linearity, and applying the weak convergence from Assumption 3. The first term is $o(1)$ because the approximations can be asymptotically replaced by the limiting mappings, i.e. because for all $\theta \in \Theta$ and all $r \in \{0, 1, \ldots, s\}$, we have

$$\begin{aligned}
&\left\| \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] \right. \\
&\left. - \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbb{E}_\theta \left[\varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right] \right\| \\
&\overset{(1)}{\le} \frac{1}{n-m} \sum_{t=m+1}^{n} \left\| \varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) - \varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(r)}(\varepsilon_t^{(r)}, \theta) \right]^\top \right) \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(2)}{\le} L_\varphi \frac{1}{n-m} \sum_{t=m+1}^{n} \left\| \mathrm{Vec}\left(G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta) \right) - \mathrm{Vec}\left(G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(r)}(\varepsilon_t^{(r)}, \theta) \right) \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(3)}{\le} L_\varphi \frac{1}{n-m} \sum_{t=m+1}^{n} \sum_{j=0}^{m} \left\| G_{t-j}^{(n,r)}(\varepsilon_{t-j}^{(r)}, \theta) - G_{t-j}^{(r)}(\varepsilon_{t-j}^{(r)}, \theta) \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(4)}{\le} L_\varphi \frac{1}{1-m/n}(m+1) \left(\frac{1}{n} \sum_{t=1}^{n} \left\| G_t^{(n,r)}(\varepsilon_0^{(r)}, \theta) - G_t^{(r)}(\varepsilon_0^{(r)}, \theta) \right\|_{\mathcal{L}^1(\theta)} \right) \\
&\overset{(5)}{=} o(1),
\end{aligned}$$

where (1) holds by linearity of summation, the triangle inequality, linearity of expectation, and Jensen's inequality, (2) holds because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), (3) holds by the triangle inequality and Minkowski inequality, (4) follows by upper bounding the original quantity by $m+1$ times full sum (i.e. by including more non-negative terms in the summation), factoring out $1/n$, and noting that each $(G_t^{(i,r)}(\varepsilon_j, \theta))_{j \in \mathbb{Z}}$, $(G_t^{(r)}(\varepsilon_j, \theta))_{j \in \mathbb{Z}}$ is a stationary ergodic process for each fixed $t \in \mathbb{N}$, $i \in \mathbb{N}$, and $r = 0, 1, \ldots, s$, which allows us to shift the noise inputs, and (5) is from applying the convergence condition from Assumption 2.

Step 2: Smoothness. Next, we show that $\mathcal{C}$ from Theorem 2 consists of C^1-smooth functions Φ. To begin, we establish that the map $\theta \mapsto \Phi(\theta)$ is differentiable on the open set O from Assumption 4. For any $\varphi \in \mathcal{F}^k$, define the linear functional

$$T_\varphi(\mu) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\mu(x),$$

which maps probability measures on $\mathbb{R}^{(m+1)\times d}$ to $\mathbb{R}^k$. For $\theta \in O$, denote the derivative

$$D\Phi(\theta)[h] = D\left(\int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\mathbf{p}(\theta)(x) \right)[h] = \int_{\mathbb{R}^{(m+1)\times d}} \varphi\, d(D\mathbf{p}(\theta)[h]),$$

where $h \in \mathbb{R}^p$. For all $\theta \in O$, we have, as $\|h\| \to 0$,

$$\begin{aligned}
&\frac{\|\Phi(\theta+h) - \Phi(\theta) - D\Phi(\theta)[h]\|}{\|h\|} \\
&\overset{(1)}{=} \frac{\left\| \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\mathbf{p}(\theta+h)(x) - \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\mathbf{p}(\theta)(x) - \int_{\mathbb{R}^{(m+1)\times d}} \varphi\, d(D\mathbf{p}(\theta)[h]) \right\|}{\|h\|} \\
&\overset{(2)}{=} \frac{\|T_\varphi(\mathbf{p}(\theta+h)) - T_\varphi(\mathbf{p}(\theta)) - T_\varphi(D\mathbf{p}(\theta)[h])\|}{\|h\|} \\
&\overset{(3)}{=} \frac{\|T_\varphi(\mathbf{p}(\theta+h) - \mathbf{p}(\theta) - D\mathbf{p}(\theta)[h])\|}{\|h\|} \\
&\overset{(4)}{\le} \|\varphi\|_\infty \frac{\|\mathbf{p}(\theta+h) - \mathbf{p}(\theta) - D\mathbf{p}(\theta)[h]\|_{\mathrm{TV}}}{\|h\|} \\
&\overset{(5)}{\le} \frac{\|\mathbf{p}(\theta+h) - \mathbf{p}(\theta) - D\mathbf{p}(\theta)[h]\|_{\mathrm{TV}}}{\|h\|} \\
&\overset{(6)}{\to} 0,
\end{aligned}$$

by (1) the definition of Φ, (2) the definition of T_φ, (3) the linearity of T_φ, (4) upper bounding by the sup norm multiplied by the total variation norm, (5) the boundedness of the random Fourier features so that $\|\varphi\|_\infty \leq 1$, and (6) because $\mathbf{p}$ is Fréchet C^1 on O with respect to the total variation norm by Assumption 4.

Next, we establish the continuity of $\theta \mapsto D\Phi(\theta)$ on O. By Assumption 4, we know that the map $\theta \mapsto D\mathbf{p}(\theta)$ is continuous on O. That is, for all $\theta \in O$ and all $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\theta' \in O$, if $\|\theta - \theta'\| < \delta$, then

$$\|D\mathbf{p}(\theta) - D\mathbf{p}(\theta')\|_{\mathrm{op}} = \sup_{\|h\|\leq 1} \|D\mathbf{p}(\theta)[h] - D\mathbf{p}(\theta')[h]\|_{\mathrm{TV}} < \epsilon.$$

Observe that, for all $\theta \in O$ and all $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\theta' \in O$, if $\|\theta - \theta'\| < \delta$, then for $\|D\Phi(\theta) - D\Phi(\theta')\|_{\mathrm{op}} = \sup_{\|h\|\leq 1} \|D\Phi(\theta)[h] - D\Phi(\theta')[h]\|$ we have

$$\begin{aligned}
&\|D\Phi(\theta) - D\Phi(\theta')\|_{\mathrm{op}} \\
&\overset{(1)}{=} \left\| \int_{\mathbb{R}^{(m+1)\times d}} \varphi\, d(D\mathbf{p}(\theta)) - \int_{\mathbb{R}^{(m+1)\times d}} \varphi\, d(D\mathbf{p}(\theta')) \right\|_{\mathrm{op}} \\
&\overset{(2)}{=} \|T_\varphi(D\mathbf{p}(\theta)) - T_\varphi(D\mathbf{p}(\theta'))\|_{\mathrm{op}} \\
&\overset{(3)}{=} \|T_\varphi(D\mathbf{p}(\theta) - D\mathbf{p}(\theta'))\|_{\mathrm{op}} \\
&\overset{(4)}{\leq} \|\varphi\|_\infty \|D\mathbf{p}(\theta) - D\mathbf{p}(\theta')\|_{\mathrm{op}} \\
&\overset{(5)}{\leq} \|D\mathbf{p}(\theta) - D\mathbf{p}(\theta')\|_{\mathrm{op}} \\
&\overset{(6)}{\leq} \epsilon,
\end{aligned}$$

by (1) the definition of $D\Phi$, (2) the definition of T_φ, (3) the linearity of T_φ, (4) upper bounding by the sup norm multiplied by the operator norm, (5) the boundedness of the random Fourier features so that $\|\varphi\|_\infty \leq 1$, and (6) because $\mathbf{p}$ is Fréchet C^1 on O with respect to the total variation norm by Assumption 4. This shows that the map $\theta \mapsto \Phi(\theta)$ is C^1 on O. The (7) extension theorem ensures the existence of a C^1 extension to the rest of $\mathbb{R}^p$. Therefore, $\mathcal{C}$ consists of C^1-smooth functions, i.e. $\mathcal{C} \subset C^1(\mathbb{R}^p, \mathbb{R}^k)$.

Step 3: Embeddings. Lastly, by Assumption 4, Θ is a compact subset of $\mathbb{R}^p$ with non-empty interior, so the box-counting dimension is p. Thus, once $k \geq 2p+1$, the fractal Whitney embedding prevalence theorem (Theorem 1) implies that embeddings are prevalent in $C^1(\Theta, \mathbb{R}^k)$; specifically, "almost every" (in the sense of prevalence) smooth map from $\mathbb{R}^p$ to $\mathbb{R}^k$ is one-to-one on Θ and is an immersion on each compact subset of a smooth manifold contained in Θ.

□

2.3. Proof of Theorem 3.

Setup. We will use Corollary 3.2 of (10) to conclude that $\hat\theta^{\mathrm{TA}}$ is consistent. The sample discrepancy function, from Eq. (6), is given by

$$\hat Q_n^{\mathrm{TA}}(\theta) = F^{\mathrm{obs}} - \bar F^{\mathrm{sim}}(\theta),$$

and the population discrepancy function is given by

$$Q^{\mathrm{TA}}(\theta) = \Phi(\theta_0) - \Phi(\theta).$$

To use Corollary 3.2 of (10), it suffices to show that, under the assumptions of Theorem 3, the following conditions hold:

1. $\left\|\hat Q_n^{\mathrm{TA}}(\hat\theta^{\mathrm{TA}})\right\| \leq o_p(1) + \inf_{\theta\in\Theta} \left\|\hat Q_n^{\mathrm{TA}}(\theta)\right\|$.
2. $\inf_{\|\theta-\theta_0\|>\delta} \left\|Q^{\mathrm{TA}}(\theta)\right\| > 0$ for each $\delta > 0$.
3. $\sup_{\theta\in\Theta} \frac{\left\|\hat Q_n^{\mathrm{TA}}(\theta) - Q^{\mathrm{TA}}(\theta)\right\|}{1 + \left\|\hat Q_n^{\mathrm{TA}}(\theta)\right\| + \|Q^{\mathrm{TA}}(\theta)\|} = o_p(1)$.

Condition 1. We defined the time-average estimator from Eq. (5) as

$$\hat\theta^{\mathrm{TA}} = \operatorname*{argmin}_{\theta\in\Theta} \left\|\hat Q_n^{\mathrm{TA}}(\theta)\right\|,$$

with a slight abuse of notation (since argmin is a set) because we have implicitly assumed that the sample objective function attains the infimum and that it is unique, i.e.

$$\left\|\hat{Q}_n^{\mathrm{TA}}(\hat{\theta}^{\mathrm{TA}})\right\| = \inf_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta)\right\|.$$

Therefore, the desired result is immediate

$$\left\|\hat{Q}_n^{\mathrm{TA}}(\hat{\theta}^{\mathrm{TA}})\right\| = \inf_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta)\right\| \le o_p(1) + \inf_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta)\right\|.$$

Condition 2. We establish that the infimum is strictly positive through contradiction. Assume for contradiction that there exists a $\delta > 0$ such that

$$\inf_{\|\theta-\theta_0\|>\delta}\left\|Q^{\mathrm{TA}}(\theta)\right\| = 0.$$

Then by applying the definition of infimum, there exists a sequence $(\theta_n)_{n\in\mathbb{N}}$, where $\|\theta_n - \theta_0\| > \delta$ for all $n \in \mathbb{N}$, such that $\lim_{n\to\infty}\left\|Q^{\mathrm{TA}}(\theta_n)\right\| = 0$ by continuity with respect to θ. By Assumption 4, Θ is a compact subset of $\mathbb{R}^p$, hence Θ is also sequentially compact, i.e. every sequence of parameters in Θ has a convergent subsequence converging to a parameter in Θ. Thus, there exists a subsequence $(\theta_{n_j})_{j\in\mathbb{N}}$, where $\left\|\theta_{n_j} - \theta_0\right\| > \delta$ for all $j \in \mathbb{N}$, and a parameter $\theta^* \in \Theta$ satisfying $\left\|Q^{\mathrm{TA}}(\theta^*)\right\| = 0$, such that $\theta_{n_j} \to \theta^*$ as $j \to \infty$. By Assumption 5, the function Φ is one-to-one, so the function Q^{TA} only takes on the value zero at the true parameter θ_0, which means that $\theta^* = \theta_0$.

However, by the continuity of $\theta \mapsto \|\theta - \theta_0\|$, we have $\left\|\theta_{n_j} - \theta_0\right\| \to \|\theta^* - \theta_0\|$ as $j \to \infty$. Therefore, $\|\theta^* - \theta_0\| \ge \delta$ because $\left\|\theta_{n_j} - \theta_0\right\| > \delta$ for all $j \in \mathbb{N}$, so we have

$$0 = \|\theta_0 - \theta_0\| = \|\theta^* - \theta_0\| \ge \delta > 0,$$

which is a contradiction.

Condition 3. It suffices to show that $\sup_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta) - Q^{\mathrm{TA}}(\theta)\right\| = o_p(1)$ because the denominator is always greater than or equal to one. For $\theta \in \Theta$ and $r \in \{0, 1, \ldots, s\}$, denote

$$F^{(r)}(\theta) = \frac{1}{n-m}\sum_{t=m+1}^{n}\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_t^{(n,r)}(\varepsilon_t^{(r)}, \theta)\right]^\top\right). \quad \textbf{[19]}$$

We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta) - Q^{\mathrm{TA}}(\theta)\right\| \\
&\overset{(1)}{\le} \left\|F^{\mathrm{obs}} - \Phi(\theta_0)\right\| + \sup_{\theta\in\Theta}\left\|\bar{F}^{\mathrm{sim}}(\theta) - \Phi(\theta)\right\| \\
&\overset{(2)}{\le} \left\|F^{\mathrm{obs}} - \Phi(\theta_0)\right\| + \frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \Phi(\theta)\right\| \\
&\overset{(3)}{\le} \left\|F^{(0)}(\theta_0) - \Phi(\theta_0)\right\| + \frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \Phi(\theta)\right\| \\
&\overset{(4)}{\le} \sup_{\theta\in\Theta}\left\|F^{(0)}(\theta) - \Phi(\theta)\right\| + \frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \Phi(\theta)\right\|,
\end{aligned}$$

by (1) the triangle inequality and subadditivity of the supremum, (2) the triangle inequality and subadditivity of the supremum, (3) noting that $F^{\mathrm{obs}} = F^{(0)}(\theta_0)$ as in Eq. (13) and Eq. (19), and (4) upper bounding by the supremum over $\theta \in \Theta$.

Observe that, by the triangle inequality, for each $r \in \{0, 1, \ldots, s\}$, we have

$$\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \Phi(\theta)\right\| \le \sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \mathbb{E}_\theta\left[F^{(r)}(\theta)\right]\right\| + \sup_{\theta\in\Theta}\left\|\mathbb{E}_\theta\left[F^{(r)}(\theta)\right] - \Phi(\theta)\right\|. \quad \textbf{[20]}$$

The next steps are to show that the first term is $o_p(1)$ and the second term is $o(1)$.

Step 1. First, we consider the term $\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \mathbb{E}_\theta\left[F^{(r)}(\theta)\right]\right\|$. Since $\Theta \subset \mathbb{R}^p$ is compact by Assumption 4, it is totally bounded, so there exists an ϵ-net, i.e. for every $\epsilon > 0$, there exist finitely many points $\theta_1, \ldots, \theta_{N(\epsilon)} \in \Theta$ for some $N(\epsilon) \in \mathbb{N}$, such that

$$\Theta \subset \bigcup_{i\in[N(\epsilon)]} B_\epsilon(\theta_i),$$

where $B_\epsilon(\theta_i)$ is a ball of radius ϵ. Thus, $\forall\theta \in \Theta$, there exists an $i \in [N(\epsilon)]$ such that $\|\theta - \theta_i\| \le \epsilon$. By adding and subtracting the same term, the triangle inequality, the definition of the ϵ-net for some $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, and the definition of the supremum, we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta} \left\| F^{(r)}(\theta) - \mathbb{E}_\theta \left[F^{(r)}(\theta) \right] \right\| \\
&\le \max_{i\in[N(\epsilon)]} \left\| F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i} \left[F^{(r)}(\theta_i) \right] \right\| \\
&+ \sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| \left(F^{(r)}(\theta) - \mathbb{E}_\theta \left[F^{(r)}(\theta) \right]\right) - \left(F^{(r)}(\theta') - \mathbb{E}_{\theta'} \left[F^{(r)}(\theta') \right]\right) \right\| .
\end{aligned}
$$

Taking the expectation $\mathbb{E}(\cdot)$ with respect to the law of the noise inputs, we have

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\theta\in\Theta} \left\| F^{(r)}(\theta) - \mathbb{E}_\theta \left[F^{(r)}(\theta) \right] \right\| \right) \\
&\overset{(1)}{\le} \mathbb{E}\left(\max_{i\in[N(\epsilon)]} \left\| F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i} \left[F^{(r)}(\theta_i) \right] \right\| \right) \\
&+ \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| \left(F^{(r)}(\theta) - \mathbb{E}_\theta \left[F^{(r)}(\theta) \right]\right) - \left(F^{(r)}(\theta') - \mathbb{E}_{\theta'} \left[F^{(r)}(\theta') \right]\right) \right\| \right) \\
&\overset{(2)}{\le} \sum_{i\in[N(\epsilon)]} \mathbb{E}_{\theta_i}\left(\left\| F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i} \left[F^{(r)}(\theta_i) \right] \right\| \right) \\
&+ \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| \left(F^{(r)}(\theta) - \mathbb{E}_\theta \left[F^{(r)}(\theta) \right]\right) - \left(F^{(r)}(\theta') - \mathbb{E}_{\theta'} \left[F^{(r)}(\theta') \right]\right) \right\| \right) \\
&\overset{(3)}{\le} \sum_{i\in[N(\epsilon)]} \mathbb{E}_{\theta_i}\left(\left\| F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i} \left[F^{(r)}(\theta_i) \right] \right\| \right) \\
&+ \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| F^{(r)}(\theta) - F^{(r)}(\theta') \right\| \right) + \sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| \mathbb{E}_\theta \left[F^{(r)}(\theta) \right] - \mathbb{E}_{\theta'} \left[F^{(r)}(\theta') \right] \right\| \\
&\overset{(4)}{\le} \sum_{i\in[N(\epsilon)]} \mathbb{E}_{\theta_i}\left(\left\| F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i} \left[F^{(r)}(\theta_i) \right] \right\| \right) \\
&+ 2\, \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\le\epsilon}} \left\| F^{(r)}(\theta) - F^{(r)}(\theta') \right\| \right),
\end{aligned}
$$

by (1) the previous ϵ-net inequality, monotonicity of expectation, and linearity of expectation, (2) upper bounding the maximum by the sum and linearity of expectation, (3) the triangle inequality and subadditivity of the supremum, and (4) linearity of expectation, Jensen's inequality, and monotonicity of expectation.

Step 1.1. We begin with the first term. For all $i \in N(\epsilon)$, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_i}\left(\left\|F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}\left[F^{(r)}(\theta_i)\right]\right\|\right) \\
&\overset{(1)}{\leq} 2\ C\ \frac{1}{n-m}\ \sum_{j=1}^{\infty}\left(\sum_{t=m+1}^{n}(\eta_{t,j,2,2}^{(n,r)}(\theta_i))^2\right)^{\frac{1}{2}} \\
&\overset{(2)}{\leq} 2\ C\ \frac{1}{n-m}\ \sum_{j=1}^{\infty}(n-m)^{\frac{1}{2}} \max_{t\in\{m+1,\ldots,n\}} \eta_{t,j,2,2}^{(n,r)}(\theta_i) \\
&\overset{(3)}{\leq} 2\ C\ \frac{1}{\sqrt{n-m}}\ \Psi^{\varphi}\ \sum_{j=1}^{\infty} j^{-\beta^{\varphi}} \\
&\overset{(4)}{\leq} 2\ C\ \frac{1}{\sqrt{n-m}}\ \Psi^{\varphi} K,
\end{aligned}
$$

by (1) factoring out $\frac{1}{n-m}$ then applying the second inequality from Lemma 3 to the time series of random features and further upper bounding with infinitely many terms in the summation, (2) upper bounding the sum by $n-m$ times the maximum and taking the square root of both terms, (3) by bounds on the physical dependence measures of the original time series by Assumption 7 and of the time series of random features by Lemma 2, and (4) because $\beta^{\varphi} > 2$ so the series converges. Hence, for each fixed $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, for all $i \in N(\epsilon)$, we have

$$
\mathbb{E}_{\theta_i}\left(\left\|F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}\left[F^{(r)}(\theta_i)\right]\right\|\right) = O(n^{-\frac{1}{2}}).
$$

Step 1.2. We now consider the second term. Observe that

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\leq\epsilon}} \left\|F^{(r)}(\theta) - F^{(r)}(\theta')\right\|\right) \\
&\overset{(1)}{\leq} L_{\varphi}\frac{1}{n-m}\sum_{t=m+1}^{n}\sum_{j=0}^{m}\mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\leq\epsilon}} \left\|G_{t-j}^{(n,r)}(\varepsilon_{t-j}^{(r)},\theta) - G_{t-j}^{(n,r)}(\varepsilon_{t-j}^{(r)},\theta')\right\|\right) \\
&\overset{(2)}{\leq} L_{\varphi}(m+1)\eta_G,
\end{aligned}
$$

because (1) the random Fourier features φ are L_{φ}-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, subadditivity of the supremum, and linearity of expectation, and (2) by upper bounding the average by the supremum over t and applying the stochastic equicontinuity-type condition from Assumption 6, noting that each mapping $(G_t^{(i,r)}(\varepsilon_j,\theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process for each fixed $t \in \mathbb{N}$, $i \in \mathbb{N}$, $r = 0, 1, \ldots, s$, which allows us to use this condition even with shifted noise inputs. By Assumption 6, $\eta_G \to 0$ as $\epsilon \to 0$. However, taking $\epsilon \to 0$ makes $N(\epsilon) \to \infty$, so we must specify the rates.

Previously, we showed that, for each fixed $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, for all $i \in N(\epsilon)$, we have

$$
\mathbb{E}_{\theta_i}\left(\left\|F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}\left[F^{(r)}(\theta_i)\right]\right\|\right) = O(n^{-\frac{1}{2}}).
$$

Therefore, for any sequence $(\epsilon_n)_{n\in\mathbb{N}}$, $\epsilon_n \to 0$ as $n \to \infty$, such that $N(\epsilon_n) = o(n^{\frac{1}{2}})$, we have

$$
\begin{aligned}
&\sum_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left(\left\|F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}\left[F^{(r)}(\theta_i)\right]\right\|\right) \\
&\leq N(\epsilon)\max_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left(\left\|F^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}\left[F^{(r)}(\theta_i)\right]\right\|\right) \\
&= o(1),
\end{aligned}
$$

and $\eta_G = \eta_G(\epsilon_n) = o(1)$ so that the upper bound from (2) is $L_{\varphi}(m+1)\eta_G = o(1)$.

Putting everything together, by Markov's inequality, we have the desired result

$$
\sup_{\theta\in\Theta}\left\|F^{(r)}(\theta) - \mathbb{E}_{\theta}\left[F^{(r)}(\theta)\right]\right\| = o_p(1).
$$

Step 2. Second, we consider the term $\sup_{\theta\in\Theta}\left\|\mathbb{E}_\theta\left[F^{(r)}(\theta)\right]-\Phi(\theta)\right\|$. For each $r\in\{0,1,\ldots,s\}$,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\mathbb{E}_\theta\left[F^{(r)}(\theta)\right]-\Phi(\theta)\right\| \\
&\overset{(1)}{=}\sup_{\theta\in\Theta}\left\|\frac{1}{n-m}\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]-\Phi(\theta)\right\| \\
&\overset{(2)}{\le}\sup_{\theta\in\Theta}\left\|\begin{matrix}\dfrac{1}{n-m}\displaystyle\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]\\ -\dfrac{1}{n-m}\displaystyle\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]\end{matrix}\right\| \\
&+\sup_{\theta\in\Theta}\left\|\frac{1}{n-m}\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]-\Phi(\theta)\right\| \\
&\overset{(3)}{=}o(1),
\end{aligned}
$$

by (1) the definition of $F^{(r)}(\theta)$ from Eq. (19) and linearity of expectation, (2) adding and subtracting the same term, the triangle inequality, and subadditivity of the supremum, (3) because both terms are $o(1)$. Specifically, the second term is $o(1)$ by upper bounding the ℓ_2 norm $\|\cdot\|$ by the ℓ_1 norm (i.e. over the k random features) using the monotonicity of norms, subadditivity of the supremum, linearity, and applying the weak convergence from Assumption 3. The first term is $o(1)$ because the approximations can be asymptotically replaced by the limiting mappings, because for all $r\in\{0,1,\ldots,s\}$, we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\begin{matrix}\dfrac{1}{n-m}\displaystyle\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]\\ -\dfrac{1}{n-m}\displaystyle\sum_{t=m+1}^{n}\mathbb{E}_\theta\left[\varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\right]\end{matrix}\right\| \\
&\overset{(1)}{\le}\sup_{\theta\in\Theta}\left(\frac{1}{n-m}\sum_{t=m+1}^{n}\left\|\begin{matrix}\varphi\left(\left[G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\\ -\varphi\left(\left[G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(r)}(\varepsilon_{t}^{(r)},\theta)\right]^\top\right)\end{matrix}\right\|_{\mathcal{L}^1(\theta)}\right) \\
&\overset{(2)}{\le}L_\varphi\sup_{\theta\in\Theta}\left(\frac{1}{n-m}\sum_{t=m+1}^{n}\left\|\begin{matrix}\mathrm{Vec}\left(G_{t-m}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right)\\ -\mathrm{Vec}\left(G_{t-m}^{(r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t}^{(r)}(\varepsilon_{t}^{(r)},\theta)\right)\end{matrix}\right\|_{\mathcal{L}^1(\theta)}\right) \\
&\overset{(3)}{\le}L_\varphi\sup_{\theta\in\Theta}\left(\frac{1}{n-m}\sum_{t=m+1}^{n}\sum_{j=0}^{m}\left\|G_{t-j}^{(n,r)}(\varepsilon_{t-j}^{(r)},\theta)-G_{t-j}^{(r)}(\varepsilon_{t-j}^{(r)},\theta)\right\|_{\mathcal{L}^1(\theta)}\right) \\
&\overset{(4)}{\le}L_\varphi\frac{1}{1-m/n}(m+1)\sup_{\theta\in\Theta}\left(\frac{1}{n}\sum_{t=1}^{n}\left\|G_{t}^{(n,r)}(\varepsilon_{0}^{(r)},\theta)-G_{t}^{(r)}(\varepsilon_{0}^{(r)},\theta)\right\|_{\mathcal{L}^1(\theta)}\right) \\
&\overset{(5)}{=}o(1),
\end{aligned}
$$

where (1) holds by linearity of summation, the triangle inequality, linearity of expectation, and Jensen's inequality, (2) holds because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), (3) holds by the triangle inequality and Minkowski inequality, (4) follows by upper bounding the original quantity by $m+1$ times full sum (i.e. by including more non-negative terms in the summation), factoring out $1/n$, and noting that each $(G_t^{(i,r)}(\varepsilon_j,\theta))_{j\in\mathbb{Z}}$, $(G_t^{(r)}(\varepsilon_j,\theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process for each fixed $t\in\mathbb{N}$, $i\in\mathbb{N}$, and $r=0,1,\ldots,s$, which allows us to shift the noise inputs, and (5) is from applying the convergence condition from Assumption 2.

Step 3. We have shown that for each $r\in\{0,1,\ldots,s\}$, the terms from Eq. (20) are $o_p(1)$ and $o(1)$, respectively. Putting it all together, we have the desired result that

$$\sup_{\theta\in\Theta}\left\|\hat{Q}_n^{\mathrm{TA}}(\theta)-Q^{\mathrm{TA}}(\theta)\right\|=o_p(1).$$

□

2.4. Proof of Theorem 4.

See the proof of Theorem 2.3 in (5).

□

2.5. Proof of Theorem 5.

Step 1: Limiting local expectations. We claim that under Assumptions 8-10, $\forall \tilde{\theta}_u \in \tilde{\Theta}_u$, the limiting local expectation values of the k random features of the time series are given by

$$\tilde{\Phi}_u(\tilde{\theta}_u) = \int_{\mathbb{R}^{(m+1)\times d}} \varphi(x)\, d\tilde{P}_{\tilde{\theta}_u,u}(x), \quad u \in [0,1].$$

That is, for all $\theta \in \Theta$, all $u \in [0,1]$, and all $r \in \{0,1,\ldots,s\}$, we have, as $n \to \infty$,

$$\left\| \mathbb{E}_\theta \left[\varphi\left(\left[G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] - \tilde{\Phi}_u(\tilde{\theta}_u) \right\| = o(1). \quad \textbf{[21]}$$

Note that the time series has this causal representation by Assumptions 8, and that we only observe the time series at rescaled times t/n for $t = 1, \ldots, n$.

The desired result from Eq. (21) follows because, for all $\theta \in \Theta$, all $u \in [0,1]$, and all $r \in \{0,1,\ldots,s\}$, we have

$$\begin{aligned}
&\left\| \mathbb{E}_\theta \left[\varphi\left(\left[G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] - \tilde{\Phi}_u(\tilde{\theta}_u) \right\| \\
&\overset{(1)}{\leq} \left\| \begin{aligned} &\mathbb{E}_\theta \left[\varphi\left(\left[G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] \\ &- \mathbb{E}_\theta \left[\varphi\left(\left[G_u(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] \end{aligned} \right\| \\
&+ \left\| \mathbb{E}_\theta \left[\varphi\left(\left[G_u(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] - \tilde{\Phi}_u(\tilde{\theta}_u) \right\| \\
&\overset{(2)}{=} o(1),
\end{aligned}$$

where (1) is from adding and subtracting the corresponding term with the limiting mapping and by applying the triangle inequality, and (2) holds because the second term is equal to zero by using the equivalent causal representation from Assumption 10 and because the first term is $o(1)$. Specifically, because the approximations can be asymptotically replaced by the limiting mappings, because for all $\theta \in \Theta$, all $u \in [0,1]$, and all $r \in \{0,1,\ldots,s\}$, we have

$$\begin{aligned}
&\left\| \begin{aligned} &\mathbb{E}_\theta \left[\varphi\left(\left[G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] \\ &- \mathbb{E}_\theta \left[\varphi\left(\left[G_u(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u(\varepsilon_0^{(r)},\theta) \right]^\top \right) \right] \end{aligned} \right\| \\
&\overset{(1)}{\leq} \left\| \begin{aligned} &\varphi\left(\left[G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right]^\top \right) \\ &- \varphi\left(\left[G_u(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u(\varepsilon_0^{(r)},\theta) \right]^\top \right) \end{aligned} \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(2)}{\leq} L_\varphi \left\| \begin{aligned} &\mathrm{Vec}\left(G_u^{(n,r)}(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u^{(n,r)}(\varepsilon_0^{(r)},\theta) \right) \\ &- \mathrm{Vec}\left(G_u(\varepsilon_{-m}^{(r)},\theta), \ldots, G_u(\varepsilon_0^{(r)},\theta) \right) \end{aligned} \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(3)}{\leq} L_\varphi \sum_{j=0}^{m} \left\| G_u^{(n,r)}(\varepsilon_{-j}^{(r)},\theta) - G_u(\varepsilon_{-j}^{(r)},\theta) \right\|_{\mathcal{L}^1(\theta)} \\
&\overset{(4)}{=} o(1),
\end{aligned}$$

where (1) holds by linearity of expectation and Jensen's inequality, (2) holds because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), (3) holds by the triangle inequality and Minkowski inequality, and (4) follows from the convergence condition from Assumption 9, noting that each mapping $(G_u^{(i,r)}(\varepsilon_j,\theta))_{j\in\mathbb{Z}}$, $(G_u(\varepsilon_j,\theta))_{j\in\mathbb{Z}}$ is a stationary ergodic process for each fixed $u \in [0,1]$, $i \in \mathbb{N}$, $r \in \{0,1,\ldots,s\}$, which allows us to use the convergence condition regardless of the shifts in the noise inputs.

Step 2: Smoothness. The arguments used to show that $\tilde{\Phi}_u$, $u \in [0,1]$, are C^1-smooth in the parameter values are analogous to those used in the proof of Theorem 2 to show that Φ is C^1-smooth in the parameter values. Specifically, by Assumption 11, $\tilde{\mathbf{p}}_u$ is Fréchet C^1 on $\tilde{O}_u$ with respect to the total variation norm (cf. Assumption 4), so the same arguments used in the proof of Theorem 2 hold when replacing θ, Φ, P_θ, $\mathcal{C}$, O, $\mathbf{p}$ with $\tilde{\theta}_u$, $\tilde{\Phi}_u$, $\tilde{P}_{\tilde{\theta}_u,u}$, $\tilde{\mathcal{C}}_u$, $\tilde{O}_u$, $\tilde{\mathbf{p}}_u$ at a particular rescaled time $u \in [0,1]$. The details are omitted due to similarity.

Step 3: Embeddings. For each $u \in [0,1]$, the box-counting dimension of $\tilde{\Theta}_u$ is less than or equal to p because: (1) Θ has box-counting dimension p as it is a compact subset of $\mathbb{R}^p$ with non-empty interior by Assumption 11, and (2) $\tilde{\Theta}_u = g_u(\Theta)$ and g_u is Lipschitz by Assumption 10 so $\dim(\tilde{\Theta}_u) = \dim(g_u(\Theta)) \leq \dim(\Theta) = p$. Therefore, once $k \geq 2p+1$, the fractal Whitney embedding prevalence theorem (Theorem 1) implies that for each $u \in [0,1]$ (i.e. applying the theorem at each u), embeddings are prevalent in $C^1(\tilde{\Theta}_u, \mathbb{R}^k)$; specifically, "almost every" (in the sense of prevalence) smooth map from $\mathbb{R}^{\tilde{p}}$ to $\mathbb{R}^k$ is one-to-one on $\tilde{\Theta}_u$ and is an immersion on each compact subset of a smooth manifold contained in $\tilde{\Theta}_u$.

□

2.6. Proof of Theorem 6.

Setup. We will use Corollary 3.2 of (10) to conclude that $\hat{\theta}^{\mathrm{RW}}$ is consistent. The sample discrepancy function, from Eq. (9), is given by

$$\hat{Q}_n^{\mathrm{RW}}(\theta) = \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left[\left\| F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right\|^2 + K_t(\theta) \right],$$

$$K_t(\theta) = 2 \left(F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right)^\top \left(\left[f_t^{\mathrm{obs}} - \bar{f}_t^{\mathrm{sim}}(\theta) \right] - \left[F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right] \right),$$

and the population discrepancy function is given by

$$Q^{\mathrm{RW}}(\theta) = \int_0^1 \left\| \Phi_u(\theta_0) - \Phi_u(\theta) \right\|^2 du.$$

To use Corollary 3.2 of (10), it suffices to show that, under the assumptions of Theorem 6, the following conditions hold:

1. $\left| \hat{Q}_n^{\mathrm{RW}}(\hat{\theta}^{\mathrm{RW}}) \right| \leq o_p(1) + \inf_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|$.
2. $\inf_{\|\theta - \theta_0\| > \delta} \left| Q^{\mathrm{RW}}(\theta) \right| > 0$ for each $\delta > 0$.
3. $\sup_{\theta \in \Theta} \frac{\left| \hat{Q}_n^{\mathrm{RW}}(\theta) - Q^{\mathrm{RW}}(\theta) \right|}{1 + \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right| + \left| Q^{\mathrm{RW}}(\theta) \right|} = o_p(1)$.

Condition 1. We defined the rolling-window estimator from Eq. (8)

$$\hat{\theta}^{\mathrm{RW}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|,$$

with a slight abuse of notation (since argmin is a set) because we have implicitly assumed that the sample objective function attains the infimum and that it is unique, i.e.

$$\left| \hat{Q}_n^{\mathrm{RW}}(\hat{\theta}^{\mathrm{RW}}) \right| = \inf_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|.$$

Therefore, the desired result is immediate

$$\left| \hat{Q}_n^{\mathrm{RW}}(\hat{\theta}^{\mathrm{RW}}) \right| = \inf_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right| \leq o_p(1) + \inf_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|.$$

Condition 2. We establish that the infimum is strictly positive through contradiction. Assume for contradiction that there exists a $\delta > 0$ such that

$$\inf_{\|\theta - \theta_0\| > \delta} \left| Q^{\mathrm{RW}}(\theta) \right| = 0.$$

Then by applying the definition of infimum, there exists a sequence $(\theta_n)_{n \in \mathbb{N}}$, where $\|\theta_n - \theta_0\| > \delta$ for all $n \in \mathbb{N}$, such that $\lim_{n \to \infty} \left| Q^{\mathrm{RW}}(\theta_n) \right| = 0$ by continuity with respect to θ and the dominated convergence theorem. By Assumption 11, Θ is a compact subset of $\mathbb{R}^p$, hence Θ is also sequentially compact, i.e. every sequence of parameters in Θ has a convergent subsequence converging to a parameter in Θ. Thus, there exists a subsequence $(\theta_{n_j})_{j \in \mathbb{N}}$, where

$\left\|\theta_{n_j} - \theta_0\right\| > \delta$ for all $j \in \mathbb{N}$, and a parameter $\theta^* \in \Theta$ satisfying $\left|Q^{\mathrm{RW}}(\theta^*)\right| = Q^{\mathrm{RW}}(\theta^*) = 0$, such that $\theta_{n_j} \to \theta^*$ as $j \to \infty$.

By Assumption 13, for all $\theta_0, \theta^* \in \Theta$, if $\theta_0 \neq \theta^*$, then there exists a subset $\mathcal{U}_{\theta_0,\theta^*} \subseteq [0,1]$ with positive Lebesgue measure such that $P_{\theta_0,u} \neq P_{\theta^*,u}$ for all $u \in \mathcal{U}_{\theta_0,\theta^*}$. Since $P_{\theta_0,u} \neq P_{\theta^*,u}$ for all $u \in \mathcal{U}_{\theta_0,\theta^*}$, we also have that $\tilde{P}_{g_u(\theta_0),u} \neq \tilde{P}_{g_u(\theta^*),u}$ for all $u \in \mathcal{U}_{\theta_0,\theta^*}$ because $\tilde{P}_{g_u(\theta),u} = P_{\theta,u}$ is the distribution of

$$[\tilde{G}_u(\varepsilon_{-m}, g_u(\theta)), \ldots, \tilde{G}_u(\varepsilon_0, g_u(\theta))]^\top = [G_u(\varepsilon_{-m}, \theta), \ldots, G_u(\varepsilon_0, \theta)]^\top.$$

Therefore, for all $u \in \mathcal{U}_{\theta_0,\theta^*}$, we have that $g_u(\theta_0) \neq g_u(\theta^*)$ because $\tilde{\theta}_u \mapsto \tilde{P}_{\tilde{\theta}_u,u}$ is a bijection for all $u \in [0,1]$ by Assumption 11. We also have that $\tilde{\Phi}_u(g_u(\theta_0)) \neq \tilde{\Phi}_u(g_u(\theta^*))$ for all $u \in \mathcal{U}_{\theta_0,\theta^*}$, because by Assumption 12, $\tilde{\theta}_u \mapsto \tilde{\Phi}_u(\tilde{\theta}_u)$ is one-to-one for all $u \in [0,1]$. This also implies that $\Phi_u(\theta_0) \neq \Phi_u(\theta^*)$ for all $u \in \mathcal{U}_{\theta_0,\theta^*}$ by the definitions of Φ_u and $\tilde{\Phi}_u$ from Eq. (16) and Eq. (17), respectively, because $\tilde{P}_{g_u(\theta),u} = P_{\theta,u}$ as discussed above. Putting it all together, we have that, if $\theta_0 \neq \theta^*$, then

$$Q^{\mathrm{RW}}(\theta^*) = \int_0^1 \left\|\Phi_u(\theta_0) - \Phi_u(\theta^*)\right\|^2 du \geq \int_{\mathcal{U}_{\theta_0,\theta^*}} \left\|\Phi_u(\theta_0) - \Phi_u(\theta^*)\right\|^2 du > 0.$$

In other words, if $Q^{\mathrm{RW}}(\theta^*) = 0$, then $\theta^* = \theta_0$.

However, by the continuity of $\theta \mapsto \|\theta - \theta_0\|$, we have $\left\|\theta_{n_j} - \theta_0\right\| \to \|\theta^* - \theta_0\|$ as $j \to \infty$. Therefore, $\|\theta^* - \theta_0\| \geq \delta$ because $\left\|\theta_{n_j} - \theta_0\right\| > \delta$ for all $j \in \mathbb{N}$, so we have

$$0 = \|\theta_0 - \theta_0\| = \|\theta^* - \theta_0\| \geq \delta > 0,$$

which is a contradiction.

Condition 3. It suffices to show that $\sup_{\theta\in\Theta} \left|\hat{Q}_n^{\mathrm{RW}}(\theta) - Q^{\mathrm{RW}}(\theta)\right| = o_p(1)$ because the denominator is always greater than or equal to one. For $\theta \in \Theta$ and $r \in \{0, 1, \ldots, s\}$, define

$$F_t^{(r)}(\theta) = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee(m+1)}^{t} \varphi\left(\left[G_{(j-m)/n}^{(n,r)}(\varepsilon_{j-m}^{(r)}, \theta), \ldots, G_{j/n}^{(n,r)}(\varepsilon_j^{(r)}, \theta)\right]^\top\right), \qquad \textbf{[22]}$$

and

$$\begin{aligned}
I_n^1(\theta) &= \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2, \\
I_n^2(\theta) &= \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left\|\mu_{\frac{t}{n}}^n(\theta) - \hat{\mu}_{t-L,n}(\theta)\right\|^2, \qquad \textbf{[23]} \\
I_n^3(\theta) &= \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} K_t^*(\theta),
\end{aligned}$$

where

$$\begin{aligned}
\mu_{\frac{t}{n}}^n(\theta) &= \mathbb{E}_{\theta_0}\left[f_t^{\mathrm{obs}}\right] - \frac{1}{s}\sum_{r=1}^{s} \mathbb{E}_\theta\left[f_t^{(r)}(\theta)\right] \\
&= \mathbb{E}_{\theta_0}\left[\varphi\left(\left[G_{(t-m)/n}^{(n,0)}(\varepsilon_{t-m}^{(0)}, \theta_0), \ldots, G_{t/n}^{(n,0)}(\varepsilon_t^{(0)}, \theta_0)\right]^\top\right)\right] \\
&\quad - \frac{1}{s}\sum_{r=1}^{s} \mathbb{E}_\theta\left[\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta)\right]^\top\right)\right], \qquad \textbf{[24]} \\
\hat{\mu}_{t,n}(\theta) &= F_t^{\mathrm{obs}} - \bar{F}_t^{\mathrm{sim}}(\theta), \\
K_t(\theta) &= 2\hat{\mu}_{t-L,n}(\theta)^\top \left(\left[f_t^{\mathrm{obs}} - \bar{f}_t^{\mathrm{sim}}(\theta)\right] - \hat{\mu}_{t-L,n}(\theta)\right), \\
K_t^*(\theta) &= 2\hat{\mu}_{t-L,n}(\theta)^\top \left(\left[f_t^{\mathrm{obs}} - \bar{f}_t^{\mathrm{sim}}(\theta)\right] - \mu_{\frac{t}{n}}^n(\theta)\right).
\end{aligned}$$

We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\hat{Q}_n^{\mathrm{RW}}(\theta)-Q^{\mathrm{RW}}(\theta)\right| \\
&\overset{(1)}{=}\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left[\left\|F_{t-L}^{\mathrm{obs}}-\bar{F}_{t-L}^{\mathrm{sim}}(\theta)\right\|^2+K_t(\theta)\right]-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right| \\
&\overset{(2)}{=}\sup_{\theta\in\Theta}\left|I_n^1(\theta)-I_n^2(\theta)+I_n^3(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right| \\
&\overset{(3)}{\le}\sup_{\theta\in\Theta}\left|I_n^1(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right|+\sup_{\theta\in\Theta}\left|I_n^2(\theta)\right|+\sup_{\theta\in\Theta}\left|I_n^3(\theta)\right|,
\end{aligned}$$

by (1) definition of $\hat{Q}_n^{\mathrm{RW}}(\theta)$ and $Q^{\mathrm{RW}}(\theta)$, (2) a Taylor expansion (see below) and the definition of $I_n^1(\theta)$, $I_n^2(\theta)$, $I_n^3(\theta)$, and (3) the triangle inequality and subadditivity of the supremum. More specifically, regarding (2), at each time t, a Taylor expansion yields

$$\left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2=\left\|\hat{\mu}_{t-L,n}(\theta)\right\|^2+2\hat{\mu}_{t-L,n}(\theta)^\top\left(\mu_{\frac{t}{n}}^n(\theta)-\hat{\mu}_{t-L,n}(\theta)\right)+\left\|\mu_{\frac{t}{n}}^n(\theta)-\hat{\mu}_{t-L,n}(\theta)\right\|^2,$$

so rearranging this equation and plugging in for $\left\|\hat{\mu}_{t-L,n}(\theta)\right\|^2=\left\|F_{t-L}^{\mathrm{obs}}-\bar{F}_{t-L}^{\mathrm{sim}}(\theta)\right\|^2$ yields

$$\begin{aligned}
&\left\|F_{t-L}^{\mathrm{obs}}-\bar{F}_{t-L}^{\mathrm{sim}}(\theta)\right\|^2+K_t(\theta) \\
&=\left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2-\left\|\mu_{\frac{t}{n}}^n(\theta)-\hat{\mu}_{t-L,n}(\theta)\right\|^2+\left[-2\hat{\mu}_{t-L,n}(\theta)^\top\left(\mu_{\frac{t}{n}}^n(\theta)-\hat{\mu}_{t-L,n}(\theta)\right)+K_t(\theta)\right] \\
&=\left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2-\left\|\mu_{\frac{t}{n}}^n(\theta)-\hat{\mu}_{t-L,n}(\theta)\right\|^2+2\hat{\mu}_{t-L,n}(\theta)^\top\left(\left[f_t^{\mathrm{obs}}-\bar{f}_t^{\mathrm{sim}}(\theta)\right]-\mu_{\frac{t}{n}}^n(\theta)\right),
\end{aligned}$$

therefore summing yields

$$\sum_{t=m+\tau+L}^{n}\left[\left\|F_{t-L}^{\mathrm{obs}}-\bar{F}_{t-L}^{\mathrm{sim}}(\theta)\right\|^2+K_t(\theta)\right]=I_n^1(\theta)-I_n^2(\theta)+I_n^3(\theta).$$

It suffices to show that

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|I_n^1(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right|&=o(1), \\
\sup_{\theta\in\Theta}\left|I_n^2(\theta)\right|&=o_p(1), \\
\sup_{\theta\in\Theta}\left|I_n^3(\theta)\right|&=o_p(1),
\end{aligned}$$

so we break down the remainder of the proof into these three steps.

Step 1. We show that

$$\sup_{\theta\in\Theta}\left|I_n^1(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right|=o(1).$$

Observe that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|I_n^1(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right| \\
&\overset{(1)}{=}\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right| \\
&\overset{(2)}{\le}\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu_{\frac{t}{n}}^n(\theta)\right\|^2-\int_0^1\left\|\mu_u^n(\theta)\right\|^2du\right| \\
&+\sup_{\theta\in\Theta}\left|\int_0^1\left\|\mu_u^n(\theta)\right\|^2du-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right|,
\end{aligned}$$

by (1) the definition of $I_n^1(\theta)$, and (2) adding and subtracting the same term, the triangle inequality, and subadditivity of the supremum. We will show that both terms are $o(1)$.

Step 1.1. Observe that

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)\right\|^2-\int_{\frac{m+\tau+L}{n}}^{1}\left\|\mu^n_u(\theta)\right\|^2du\right| \\
&\overset{(1)}{\leq}\frac{mk}{n}+\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)\right\|^2-\int_{\frac{m+\tau+L}{n}}^{1}\left\|\mu^n_u(\theta)\right\|^2du\right| \\
&\overset{(2)}{\leq}\frac{mk}{n}+\sup_{\theta\in\Theta}\int_{\frac{m+\tau+L}{n}}^{1}\left|\left\|\mu^n_{\frac{\lceil un\rceil}{n}}(\theta)\right\|^2-\left\|\mu^n_u(\theta)\right\|^2\right|du \\
&\overset{(3)}{\leq}\frac{mk}{n}+\frac{1}{n}\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\sup_{u\in\left[\frac{t-1}{n},\frac{t}{n}\right]}\left|\left\|\mu^n_{\frac{t}{n}}(\theta)\right\|^2-\left\|\mu^n_u(\theta)\right\|^2\right| \\
&\overset{(4)}{\leq}\frac{mk}{n}+\frac{1}{n}\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\sup_{u\in\left[\frac{t-1}{n},\frac{t}{n}\right]}\left(\left\|\mu^n_{\frac{t}{n}}(\theta)\right\|+\left\|\mu^n_u(\theta)\right\|\right)\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_u(\theta)\right\| \\
&\overset{(5)}{\leq}\frac{mk}{n}+4\sqrt{k}\frac{1}{n}\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\sup_{u\in\left[\frac{t-1}{n},\frac{t}{n}\right]}\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_u(\theta)\right\| \\
&\overset{(6)}{\leq}\frac{mk}{n}+4\sqrt{k}L_\varphi(m+1)\frac{1}{n}\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\sup_{u\in\left[\frac{t-1}{n},\frac{t}{n}\right]}\left\|G^{(n,r)}_{\frac{t}{n}}(\varepsilon_0^{(r)},\theta)-G^{(n,r)}_{u}(\varepsilon_0^{(r)},\theta)\right\|_{\mathcal{L}^1(\theta)} \\
&+4\sqrt{k}L_\varphi(m+1)\frac{1}{n}\frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\left(\sum_{t=m+\tau+L}^{n}\sup_{u\in\left[\frac{t-1}{n},\frac{t}{n}\right]}\left\|G^{(n,r)}_{\frac{t}{n}}(\varepsilon_0^{(r)},\theta)-G^{(n,r)}_{u}(\varepsilon_0^{(r)},\theta)\right\|_{\mathcal{L}^1(\theta)}\right) \\
&\overset{(7)}{\leq}\frac{mk}{n}+8\sqrt{k}L_\varphi(m+1)\frac{n^{1-1/\kappa}}{n}\Lambda,
\end{aligned}
$$

by (1) adding and subtracting the same term (i.e. with a factor of $\frac{1}{n}$), the triangle inequality, subadditivity of the supremum, factoring out the sum of squared norms, the equality $\frac{1}{n-m}-\frac{1}{n}=\frac{m}{n(n-m)}$, upper bounding the sum by $n-m$ times the maximum over time t of the squared norm, noting that the random Fourier features are bounded between -1 and 1 so the corresponding expectations must be as well, then upper bounding the squared norm by the number of random features k, (2) the definition of the ceiling function, using the Riemann sum approximation of the integral, and the triangle inequality for integrals, (3) upper bounding the integral over each interval by the supremum over that interval multiplied by the interval length and summing over the intervals, (4) the equality $\left|x^2-y^2\right|=|x-y|\,|x+y|$ applied to the norms and the triangle inequality applied to this absolute value of the sum of norms, (5) using the definition of $\mu^n_{\frac{t}{n}}(\theta)$ and $\mu^n_u(\theta)$ and applying the triangle inequality to each of their norms, applying the triangle inequality and absolute homogeneity to the simulation-average, upper bounding each norm by $\sqrt{k}$ where $k=2p+1$, $\theta\in\mathbb{R}^p$, because each of the k dimensions of the random Fourier features takes values in $[-1,1]$, so we can upper bound by four times this amount, and the distributive property, (6) the triangle inequality, subadditivity of the supremum, linearity of expectation, Jensen's inequality, the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, the Minkowski inequality, shifting the noise inputs which is equivalent due to stationarity, and (7) Hölder's inequality applied to each sum over time t (i.e. for each $r=0,1,\dots,s$), upper bounding by the κ-variation, and applying the upper bound from Assumption 16.

Lastly, observe that

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^{n}_{\frac{t}{n}}(\theta)\right\|^{2}-\int_{0}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du\right| \\
&\overset{(1)}{\leq}\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^{n}_{\frac{t}{n}}(\theta)\right\|^{2}-\int_{\frac{m+\tau+L}{n}}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du\right| \\
&+\sup_{\theta\in\Theta}\left|\int_{\frac{m+\tau+L}{n}}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du-\int_{0}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du\right| \\
&\overset{(2)}{=}\sup_{\theta\in\Theta}\left|\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^{n}_{\frac{t}{n}}(\theta)\right\|^{2}-\int_{\frac{m+\tau+L}{n}}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du\right| \\
&+\sup_{\theta\in\Theta}\left|\int_{0}^{\frac{m+\tau+L}{n}}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du\right| \\
&\overset{(3)}{\leq}\left(\frac{mk}{n}+8\sqrt{k}L_{\varphi}(m+1)\frac{n^{1-1/\kappa}}{n}\Lambda\right)+\frac{m+\tau+L}{n}\;k \\
&\overset{(4)}{=}o(1),
\end{aligned}
$$

by (1) adding and subtracting the same term, the triangle inequality, and subadditivity of the supremum, (2) subtraction, (3) upper bounding the first term by applying the inequality derived previously, and upper bounding the second term by the length of the interval $[0,\frac{m+\tau+L}{n}]$ multiplied by k because the integrand can be upper bounded by k since the k random Fourier features each take values in $[-1,1]$, and (4) noting that both of these terms converge to zero because the first term is $o(1)$ because $\kappa\in[1,4)$ by Assumption 16 and the second term is $o(1)$ because $\tau_n=o(n)$ by Assumption 17.

Step 1.2. Observe that

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\int_{0}^{1}\left\|\mu^{n}_{u}(\theta)\right\|^{2}du-\int_{0}^{1}\left\|\Phi_{u}(\theta_{0})-\Phi_{u}(\theta)\right\|^{2}du\right| \\
&\overset{(1)}{\leq}\sup_{\theta\in\Theta}\int_{0}^{1}\left|\left\|\mu^{n}_{u}(\theta)\right\|^{2}-\left\|\Phi_{u}(\theta_{0})-\Phi_{u}(\theta)\right\|^{2}\right|du \\
&\overset{(2)}{\leq}4\sqrt{k}\sup_{\theta\in\Theta}\int_{0}^{1}\left\|\mu^{n}_{u}(\theta)-(\Phi_{u}(\theta_{0})-\Phi_{u}(\theta))\right\|du \\
&\overset{(3)}{\leq}4\sqrt{k}\int_{0}^{1}\left\|\mathbb{E}_{\theta_{0}}\left[\varphi\left(\left[G^{(n,0)}_{u}(\varepsilon^{(0)}_{-m},\theta_{0}),\ldots,G^{(n,0)}_{u}(\varepsilon^{(0)}_{0},\theta_{0})\right]^{\top}\right)\right]-\Phi_{u}(\theta_{0})\right\|du \\
&+4\sqrt{k}\frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\int_{0}^{1}\left\|\mathbb{E}_{\theta}\left[\varphi\left(\left[G^{(n,r)}_{u}(\varepsilon^{(r)}_{-m},\theta),\ldots,G^{(n,r)}_{u}(\varepsilon^{(r)}_{0},\theta)\right]^{\top}\right)\right]-\Phi_{u}(\theta)\right\|du \\
&\overset{(4)}{\leq}4\sqrt{k}\sup_{u\in[0,1]}\left\|\mathbb{E}_{\theta_{0}}\left[\varphi\left(\left[G^{(n,0)}_{u}(\varepsilon^{(0)}_{-m},\theta_{0}),\ldots,G^{(n,0)}_{u}(\varepsilon^{(0)}_{0},\theta_{0})\right]^{\top}\right)\right]-\Phi_{u}(\theta_{0})\right\| \\
&+4\sqrt{k}\frac{1}{s}\sum_{r=1}^{s}\sup_{\theta\in\Theta}\sup_{u\in[0,1]}\left\|\mathbb{E}_{\theta}\left[\varphi\left(\left[G^{(n,r)}_{u}(\varepsilon^{(r)}_{-m},\theta),\ldots,G^{(n,r)}_{u}(\varepsilon^{(r)}_{0},\theta)\right]^{\top}\right)\right]-\Phi_{u}(\theta)\right\| \\
&\overset{(5)}{=}o(1),
\end{aligned}
$$

by (1) the linearity of integrals and the triangle inequality for integrals, (2) the equality $\left|x^2-y^2\right|=|x-y|\,|x+y|$ applied to the norms and the triangle inequality applied to this absolute value of the sum of norms, then using the definition of $\mu^{n}_{u}(\theta)$ and applying the triangle inequality to each of the norms, applying the triangle inequality and absolute homogeneity to the simulation-average terms, upper bounding each norm by $\sqrt{k}$ where $k=2p+1$, $\theta\in\mathbb{R}^p$, because each of the k dimensions of the random Fourier features takes values in $[-1,1]$, so we can upper bound by four times this amount, and the distributive property, (3) the triangle inequality, subadditivity, and the distributive property, (4) upper bounding the integral by the length of the interval multiplied by the supremum over the interval, and (5) the same (u,θ)-pointwise inequalities used to show Eq. (21), where $\tilde{\Phi}_u(\tilde{\theta}_u)=\Phi_u(\theta)$ by the definitions and Assumption 10, hold uniformly over $u\in[0,1]$ and over $\theta\in\Theta$ by using the subadditivity of the supremum and because Assumption 9 is a (u,θ)-uniform convergence condition.

Putting it all together, we have

$$\sup_{\theta\in\Theta}\left|I_n^1(\theta)-\int_0^1\left\|\Phi_u(\theta_0)-\Phi_u(\theta)\right\|^2du\right|=o(1).$$

Step 2. We show that

$$\sup_{\theta\in\Theta}\left|I_n^2(\theta)\right|=o_p(1).$$

Observe that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|I_n^2(\theta)\right|\\
&\overset{(1)}{=}\sup_{\theta\in\Theta}\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)-\hat\mu_{t-L,n}(\theta)\right\|^2\\
&\overset{(2)}{\le}\sup_{\theta\in\Theta}\frac{1}{n-m}\sum_{t=m+\tau+L}^{n}\left(\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta)\right\|+\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\hat\mu_{t-L,n}(\theta)\right\|\right)^2\\
&\overset{(3)}{\le}\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta)\right\|^2\\
&+\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\hat\mu_{t-L,n}(\theta)\right\|^2,
\end{aligned}$$

by (1) the definition of $I_n^2(\theta)$, (2) adding and subtracting the same term then applying the triangle inequality, and (3) expanding the square then applying Young's inequality and subadditivity of the supremum. We will show the first term is $o(1)$ and the second term is $o_p(1)$.

Step 2.1. To begin, observe that, for every $\theta\in\Theta$, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta)\right\|^2\\
&\overset{(1)}{\le}\sum_{t=m+\tau+L}^{n}\left\|\begin{matrix}\mathbb{E}_{\theta_0}\varphi\left(\left[G^{(n,0)}_{(t-m)/n}(\varepsilon^{(0)}_{-m},\theta_0),\dots,G^{(n,0)}_{t/n}(\varepsilon^{(0)}_{0},\theta_0)\right]^\top\right)\\-\mathbb{E}_{\theta_0}\varphi\left(\left[G^{(n,0)}_{(t-L-m)/n}(\varepsilon^{(0)}_{-m},\theta_0),\dots,G^{(n,0)}_{(t-L)/n}(\varepsilon^{(0)}_{0},\theta_0)\right]^\top\right)\end{matrix}\right\|^2\\
&+\frac{1}{s}\sum_{r=1}^{s}\sum_{t=m+\tau+L}^{n}\left\|\begin{matrix}\mathbb{E}_{\theta}\varphi\left(\left[G^{(n,r)}_{(t-m)/n}(\varepsilon^{(r)}_{-m},\theta),\dots,G^{(n,r)}_{t/n}(\varepsilon^{(r)}_{0},\theta)\right]^\top\right)\\-\mathbb{E}_{\theta}\varphi\left(\left[G^{(n,r)}_{(t-L-m)/n}(\varepsilon^{(r)}_{-m},\theta),\dots,G^{(n,r)}_{(t-L)/n}(\varepsilon^{(r)}_{0},\theta)\right]^\top\right)\end{matrix}\right\|^2\\
&\overset{(2)}{\le}L^2\sum_{t=m+\tau+L}^{n}\left\|\begin{matrix}\mathbb{E}_{\theta_0}\varphi\left(\left[G^{(n,0)}_{(t-m+1)/n}(\varepsilon^{(0)}_{-m},\theta_0),\dots,G^{(n,0)}_{(t+1)/n}(\varepsilon^{(0)}_{0},\theta_0)\right]^\top\right)\\-\mathbb{E}_{\theta_0}\varphi\left(\left[G^{(n,0)}_{(t-m)/n}(\varepsilon^{(0)}_{-m},\theta_0),\dots,G^{(n,0)}_{t/n}(\varepsilon^{(0)}_{0},\theta_0)\right]^\top\right)\end{matrix}\right\|^2\\
&+\frac{L^2}{s}\sum_{r=1}^{s}\sum_{t=m+\tau+L}^{n}\left\|\begin{matrix}\mathbb{E}_{\theta}\varphi\left(\left[G^{(n,r)}_{(t-m+1)/n}(\varepsilon^{(r)}_{-m},\theta),\dots,G^{(n,r)}_{(t+1)/n}(\varepsilon^{(r)}_{0},\theta)\right]^\top\right)\\-\mathbb{E}_{\theta}\varphi\left(\left[G^{(n,r)}_{(t-m)/n}(\varepsilon^{(r)}_{-m},\theta),\dots,G^{(n,r)}_{t/n}(\varepsilon^{(r)}_{0},\theta)\right]^\top\right)\end{matrix}\right\|^2\\
&\overset{(3)}{\le}L^2L_\varphi^2(m+1)^2\sum_{t=1}^{n}\left\|G^{(n,0)}_{t/n}(\varepsilon^{(0)}_{0},\theta_0)-G^{(n,0)}_{(t-1)/n}(\varepsilon^{(0)}_{0},\theta_0)\right\|^2_{\mathcal{L}^q(\theta)}\\
&+L^2L_\varphi^2(m+1)^2\frac{1}{s}\sum_{r=1}^{s}\sum_{t=1}^{n}\left\|G^{(n,r)}_{t/n}(\varepsilon^{(r)}_{0},\theta)-G^{(n,r)}_{(t-1)/n}(\varepsilon^{(r)}_{0},\theta)\right\|^2_{\mathcal{L}^q(\theta)},
\end{aligned}$$

by (1) the definition of $\mu^n_{\frac{t}{n}}(\theta)$ and $\mu^n_{\frac{t-L}{n}}(\theta)$, the triangle inequality, expanding the square, Young's inequality, and the triangle inequality, (2) adding and subtracting the same intermediate lag terms between t and $t-L$ (i.e. $t-$

$1,\dots,t-L+1$), the triangle inequality, applying Young's inequality L times, factoring out L, adding extra lag terms so the every lag term is repeated L terms and then factoring out this second L factor, and (3) linearity of expectation, Jensen's inequality, because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, the Minkowski inequality, and Young's inequality.

Observe that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta)\right\|^2\\
&\overset{(1)}{\leq}\frac{4L^2L_\varphi^2(m+1)^2}{n-m}\sum_{r=0}^{s}\sup_{\theta\in\Theta}\sum_{t=1}^{n}\left\|G^{(n,r)}_{t/n}(\varepsilon_0^{(r)},\theta)-G^{(n,r)}_{(t-1)/n}(\varepsilon_0^{(r)},\theta)\right\|^2_{\mathcal{L}^q(\theta)}\\
&\overset{(2)}{\leq}\frac{4L^2L_\varphi^2(m+1)^2}{n-m}(s+1)n^{\max(1-2/\kappa,0)}\Lambda^2\\
&\overset{(3)}{=}o(1),
\end{aligned}$$

by (1) the previous arguments, linearity, and subadditivity of the supremum, (2) Hölder's inequality applied to each sum over time t (i.e. for each $r=0,1,\dots,s$) so we can upper bound by the squared κ-variation multiplied by $n^{\max(1-2/\kappa,0)}$, where the max is because we consider the cases $\kappa\in[1,2]$ and $\kappa\in(2,4)$ separately, and applying the upper bound $\Lambda>0$ from Assumption 16, and (3) the rate of growth of L from Assumption 17.

Step 2.2. Observe that, for every $\theta\in\Theta$, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\hat{\mu}_{t-L,n}(\theta)\right\|^2\\
&\overset{(1)}{\leq}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_{\theta_0}\varphi\left(\left[G^{(n,0)}_{(t-L-m)/n}(\varepsilon^{(0)}_{t-L-m},\theta_0),\dots,G^{(n,0)}_{(t-L)/n}(\varepsilon^{(0)}_{t-L},\theta_0)\right]^\top\right)-F^{\text{obs}}_{t-L}\right\|^2\\
&+\sum_{r=1}^{s}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_{\theta}\varphi\left(\left[G^{(n,r)}_{(t-L-m)/n}(\varepsilon^{(r)}_{t-L-m},\theta),\dots,G^{(n,r)}_{(t-L)/n}(\varepsilon^{(r)}_{t-L},\theta)\right]^\top\right)-F^{(r)}_{t-L}(\theta)\right\|^2,
\end{aligned}\tag{25}$$

by (1) the definitions of $\mu^n_{\frac{t-L}{n}}(\theta)$ and $\hat{\mu}_{t-L,n}(\theta)$, the triangle inequality, and Young's inequality, where $F_t^{(r)}(\theta)$ is from Eq. (22). We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\hat{\mu}_{t-L,n}(\theta)\right\|^2\\
&\overset{(1)}{\leq}\sum_{r=0}^{s}\sup_{\theta\in\Theta}\frac{4}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_\theta\left[f^{(r)}_{t-L}(\theta)\right]-F^{(r)}_{t-L}(\theta)\right\|^2,
\end{aligned}$$

by (1) the previous arguments, the definition of the random features $f^{(r)}_{t-L}(\theta)$ at time $t-L$ in terms of the generative mappings, linearity, and subadditivity of the supremum. It suffices to show that, for all $r=0,1,\dots,s$, we have

$$\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_\theta\left[f^{(r)}_{t-L}(\theta)\right]-F^{(r)}_{t-L}(\theta)\right\|^2=o_p(n).$$

By adding and subtracting the same term, the triangle inequality, Young's inequality, and subadditivity of the supremum, half of the previous term is less than or equal to

$$\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_\theta\left[f^{(r)}_{t-L}(\theta)\right]-\mathbb{E}_\theta F^{(r)}_{t-L}(\theta)\right\|^2+\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_\theta F^{(r)}_{t-L}(\theta)-F^{(r)}_{t-L}(\theta)\right\|^2,$$

so it suffices to show that the first term is $o(n)$ and the second term is $o_p(n)$.

Step 2.2.1. For the following, denote $\mathcal{U}_t = \left[\frac{(t-L-w)\vee(m+1)}{n}, \frac{t-L}{n}\right]$. Observe that, for all $r = 0, 1, \ldots, s$, we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta} \sum_{t=m+\tau+L}^{n} \left\| \mathbb{E}_\theta \left[f_{t-L}^{(r)}(\theta) \right] - \mathbb{E}_\theta F_{t-L}^{(r)}(\theta) \right\|^2 \\
&\overset{(1)}{\leq} \sup_{\theta\in\Theta} \sum_{t=m+\tau+L}^{n} \sup_{u\in\mathcal{U}_t} \left\| \begin{aligned} &\mathbb{E}_\theta \varphi \left(\left[G_{(t-L-m)/n}^{(n,r)}(\varepsilon_{t-L-m}^{(r)}, \theta), \ldots, G_{(t-L)/n}^{(n,r)}(\varepsilon_{t-L}^{(r)}, \theta) \right]^\top \right) \\ &- \mathbb{E}_\theta \varphi \left(\left[G_{u-m/n}^{(n,r)}(\varepsilon_{-m}^{(r)}, \theta), \ldots, G_{u}^{(n,r)}(\varepsilon_{0}^{(r)}, \theta) \right]^\top \right) \end{aligned} \right\|^2 \\
&\overset{(2)}{\leq} L_\varphi^2 (m+1)^2 \sup_{\theta\in\Theta} \sum_{t=m+\tau+L}^{n} \sup_{u\in\mathcal{U}_t} \left\| G_{(t-L)/n}^{(n,r)}(\varepsilon_0^{(r)}, \theta) - G_u^{(n,r)}(\varepsilon_0^{(r)}, \theta) \right\|_{\mathcal{L}^q(\theta)}^2 \\
&\overset{(3)}{\leq} w L_\varphi^2 (m+1)^2 n^{\max(1-2/\kappa, 0)} \left(\sup_{\theta\in\Theta} \left\| \left(G_u^{(n,r)}(\varepsilon_0, \theta) \right)_u \right\|_{\kappa\text{-var}, \mathcal{L}^q(\theta)} \right)^2 \\
&\overset{(4)}{\leq} w L_\varphi^2 (m+1)^2 n^{\max(1-2/\kappa, 0)} \Lambda^2 \\
&\overset{(5)}{=} o(n),
\end{aligned}
$$

by (1) the definitions of $F_{t-L}^{(r)}(\theta)$ and $f_{t-L}^{(r)}(\theta)$ from Eq. (22) and Eq. (24), linearity, the triangle inequality, and upper bounding the rolling-window average of the distances between the expectations by the supremum of the distance between the expectations over the corresponding rescaled time interval $\mathcal{U}_t = \left[\frac{(t-L-w)\vee(m+1)}{n}, \frac{t-L}{n}\right]$, (2) linearity of expectation, Jensen's inequality, because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, the Minkowski inequality, and Young's inequality, (3) upper bounding by w multiplied by the squared κ-variation multiplied by $n^{\max(1-2/\kappa,0)}$, where the max is because we consider the cases $\kappa \in [1, 2]$ and $\kappa \in (2, 4)$ separately, and where the w factor is from the following construction: each interval $\mathcal{U}_t$ has length at most w/n, so we consider a family of w subsequences of intervals of the form $\mathcal{U}_{j+w}, \mathcal{U}_{j+2w}, \ldots$ so they are disjoint (besides the endpoints), then upper bound the sum over each subsequence by the squared κ-variation multiplied by $n^{\max(1-2/\kappa,0)}$, then sum over all of the subsequences, which leads to the desired upper bound with the factor of w, (4) using the bound from Assumption 16, and (5) because of the rates specified in Assumption 17.

Step 2.2.2. By Markov's inequality, it suffices to show that, for all $r = 0, 1, \ldots, s$,

$$
\mathbb{E}\left(\sup_{\theta\in\Theta} \sum_{t=m+\tau+L}^{n} \left\| \mathbb{E}_\theta F_{t-L}^{(r)}(\theta) - F_{t-L}^{(r)}(\theta) \right\|^2 \right) = o(n),
$$

where $\mathbb{E}(\cdot)$ denotes expectation with respect to the law of the noise inputs. By Assumption 11, $\Theta \subset \mathbb{R}^p$ is compact, so it is totally bounded and there exists an ϵ-net, i.e. for every $\epsilon > 0$, there exist finitely many points $\theta_1, \ldots, \theta_{N(\epsilon)}$ for some $N(\epsilon) \in \mathbb{N}$, such that

$$
\Theta \subset \bigcup_{i\in[N(\epsilon)]} B_\epsilon(\theta_i),
$$

where $B_\epsilon(\theta_i)$ is a ball of radius ϵ. By adding and subtracting the same term, the triangle inequality, Young's inequality, the definition of the ϵ-net for some $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, and definition of the supremum,

$$
\begin{aligned}
&\frac{1}{2} \sup_{\theta\in\Theta} \sum_{t=m+\tau+L}^{n} \left\| \mathbb{E}_\theta \left[F_{t-L}^{(r)}(\theta) \right] - F_{t-L}^{(r)}(\theta) \right\|^2 \\
&\leq \max_{i\in[N(\epsilon)]} \sum_{t=m+\tau+L}^{n} \left\| \mathbb{E}_{\theta_i} \left[F_{t-L}^{(r)}(\theta_i) \right] - F_{t-L}^{(r)}(\theta_i) \right\|^2 \\
&+ \sup_{\substack{\theta,\theta'\in\Theta \\ \|\theta-\theta'\|\leq\epsilon}} \sum_{t=m+\tau+L}^{n} \left\| \left[\mathbb{E}_\theta \left[F_{t-L}^{(r)}(\theta) \right] - F_{t-L}^{(r)}(\theta) \right] - \left[\mathbb{E}_{\theta'} \left[F_{t-L}^{(r)}(\theta') \right] - F_{t-L}^{(r)}(\theta') \right] \right\|^2.
\end{aligned}
$$

Taking the expectation $\mathbb{E}(\cdot)$ with respect to the law of the noise inputs, we have

$$
\begin{aligned}
&\frac{1}{2}\,\mathbb{E}\left(\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_{\theta}\left[F_{t-L}^{(r)}(\theta)\right]-F_{t-L}^{(r)}(\theta)\right\|^{2}\right)\\
&\overset{(1)}{\leq}\mathbb{E}\left(\max_{i\in[N(\epsilon)]}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^{2}\right)\\
&+\mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|[\mathbb{E}_{\theta}\left[F_{t-L}^{(r)}(\theta)\right]-F_{t-L}^{(r)}(\theta)]-[\mathbb{E}_{\theta'}\left[F_{t-L}^{(r)}(\theta')\right]-F_{t-L}^{(r)}(\theta')]\right\|^{2}\right)\\
&\overset{(2)}{\leq}\sum_{i\in[N(\epsilon)]}\sum_{t=m+\tau+L}^{n}\mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^{2}\right)\\
&+\mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|[\mathbb{E}_{\theta}\left[F_{t-L}^{(r)}(\theta)\right]-F_{t-L}^{(r)}(\theta)]-[\mathbb{E}_{\theta'}\left[F_{t-L}^{(r)}(\theta')\right]-F_{t-L}^{(r)}(\theta')]\right\|^{2}\right)\\
&\overset{(3)}{\leq}\sum_{i\in[N(\epsilon)]}\sum_{t=m+\tau+L}^{n}\mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^{2}\right)\\
&+4\sum_{t=m+\tau+L}^{n}\mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}}\left\|F_{t-L}^{(r)}(\theta)-F_{t-L}^{(r)}(\theta')\right\|^{2}\right),
\end{aligned}
$$

by (1) the previous ϵ-net inequality, monotonicity of expectation, and linearity of expectation, (2) upper bounding the maximum by the sum and linearity of expectation, and (3) the triangle inequality, Young's inequality, subadditivity of the supremum, linearity of expectation, Jensen's inequality, and monotonicity of expectation.

Step 2.2.2.1. We begin with the first term. For the following derivations, define

$$\mathcal{T}_t=\{(t-L-w)\vee(m+1),\ldots,t-L-1,t-L\}.$$

For all $i\in N(\epsilon)$, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^{2}\right)\\
&\overset{(1)}{=}\operatorname{tr}\left[\operatorname{Cov}_{\theta_i}\left(F_{t-L}^{(r)}(\theta_i)\right)\right]\\
&\overset{(2)}{=}\operatorname{tr}\left[\operatorname{Var}_{\theta_i}\left(\frac{1}{w\wedge t}\sum_{j=(t-L-w)\vee(m+1)}^{t-L}\varphi\left(\left[G_{(j-m)/n}^{(n,r)}(\varepsilon_{j-m}^{(r)},\theta_i),\ldots,G_{j/n}^{(n,r)}(\varepsilon_{j}^{(r)},\theta_i)\right]^{\top}\right)\right)\right]\\
&\overset{(3)}{=}\frac{1}{(w\wedge t)^2}\sum_{\ell,j\in\mathcal{T}_t}\operatorname{tr}\left[\operatorname{Cov}_{\theta_i}\begin{pmatrix}\varphi\left(\left[G_{(\ell-m)/n}^{(n,r)}(\varepsilon_{\ell-m}^{(r)},\theta_i),\ldots,G_{\ell/n}^{(n,r)}(\varepsilon_{\ell}^{(r)},\theta_i)\right]^{\top}\right),\\ \varphi\left(\left[G_{(j-m)/n}^{(n,r)}(\varepsilon_{j-m}^{(r)},\theta_i),\ldots,G_{j/n}^{(n,r)}(\varepsilon_{j}^{(r)},\theta_i)\right]^{\top}\right)\end{pmatrix}\right]\\
&\overset{(4)}{\leq}\frac{1}{(w\wedge t)^2}\sum_{\ell,j\in\mathcal{T}_t}C^{\varphi}(|\ell-j|+1)^{-K^{\varphi}}\\
&\overset{(5)}{\leq}\frac{1}{w\wedge t}\sum_{h=-\infty}^{\infty}C^{\varphi}(|h|+1)^{-K^{\varphi}},
\end{aligned}
$$

by (1) equality of inner product with trace of outer product, and the definition of the covariance of this quantity as this outer product, (2) definition of $F_{t-L}^{(r)}(\theta_i)$ from Eq. (22), (3) expanding the trace of the covariance of the sum of random vectors by using the linearity of covariance and trace, (4) upper bounding the trace of the covariance by its trace norm and applying the trace norm inequality with constants $C^{\varphi}>0$, $K^{\varphi}>1$ from Lemma 6, (5) by enlarging the inner summation to be over all integers and upper bounding by this quantity multiplied by the number of terms in the outer summation, noting that the series converges because $K^{\varphi}>1$ by Lemma 6.

Denoting $\sum_{h=-\infty}^{\infty} C^{\varphi}(|h|+1)^{-K^{\varphi}}$ by the constant $C'>0$, since the series converges, summing over t yields

$$\begin{aligned}
&\sum_{t=m+\tau+L}^{n} \mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^2\right)\\
&\overset{(1)}{\leq} C' \sum_{t=m+\tau+L}^{n} \frac{1}{w \wedge t}\\
&\overset{(2)}{\leq} C' \frac{n-w}{w} + C' \sum_{t=m+\tau+L}^{w} \frac{1}{t}\\
&\overset{(3)}{\leq} C' \frac{n}{w} + C'(1+\log(w))\\
&\overset{(4)}{=} O(\max(n/w, \log(w))),
\end{aligned}$$

by (1) applying the previous inequality, (2) expanding the sum with the minimum, where the inequality covers both cases when w is less than or greater than $m+\tau+L$, (3) upper bounding the harmonic series and dropping the subtraction by w, (4) because $w_n = o(n)$ by Assumption 17.

Step 2.2.2.2. We now consider the second term. Observe that

$$\begin{aligned}
&\sum_{t=m+\tau+L}^{n} \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}} \left\|F_{t-L}^{(r)}(\theta)-F_{t-L}^{(r)}(\theta')\right\|^2\right)\\
&\overset{(1)}{\leq} 2\sum_{t=m+\tau}^{n} \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}} \left\| \begin{array}{l}\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta),\ldots,G_{t/n}^{(n,r)}(\varepsilon_{t}^{(r)},\theta)\right]^{\top}\right)\\ -\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)},\theta'),\ldots,G_{t/n}^{(n,r)}(\varepsilon_{t}^{(r)},\theta')\right]^{\top}\right)\end{array}\right\|^2\right)\\
&\overset{(2)}{\leq} 2(m+1)L_{\varphi}^2 \sum_{t=m+\tau}^{n}\sum_{j=0}^{m} \mathbb{E}\left(\sup_{\substack{\theta,\theta'\in\Theta\\ \|\theta-\theta'\|\leq\epsilon}} \left\|G_{(t-j)/n}^{(n,r)}(\varepsilon_{t-j}^{(r)},\theta)-G_{(t-j)/n}^{(n,r)}(\varepsilon_{t-j}^{(r)},\theta')\right\|^2\right)\\
&\overset{(3)}{\leq} 2n(m+1)^2 L_{\varphi}^2 \eta_G,
\end{aligned}$$

because (1) linearity of summation (i.e. for the rolling windows), the triangle inequality, Young's inequality, subadditivity of the supremum, linearity of expectation, and dividing by w from the rolling-window averages to cancel out the factor of w due to repeated terms, (2) the random Fourier features φ are L_{φ}-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, Young's inequality $m+1$ times, subadditivity of the supremum, and linearity of expectation, (3) upper bounding the sum by the n times the supremum over t and applying the stochastic equicontinuity-type condition from Assumption 14. By Assumption 14, $\eta_G \to 0$ as $\epsilon \to 0$. However, taking $\epsilon \to 0$ makes $N(\epsilon)\to\infty$, so we must specify the rates.

Previously, in Step 2.2.2.1 we showed that, for each fixed $\epsilon>0$ and $N(\epsilon)\in\mathbb{N}$, for all $i \in N(\epsilon)$, we have

$$\sum_{t=m+\tau+L}^{n} \mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^2\right) = O(\max(n/w,\log(w))).$$

Therefore, for any sequence $(\epsilon_n)_{n\in\mathbb{N}}$, $\epsilon_n \to 0$ as $n\to\infty$, such that $N(\epsilon_n) = o(\frac{n}{\max(n/w,\log(w))})$, we have

$$\begin{aligned}
&\sum_{i\in[N(\epsilon)]}\sum_{t=m+\tau+L}^{n} \mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^2\right)\\
&\leq N(\epsilon)\max_{i\in[N(\epsilon)]}\sum_{t=m+\tau+L}^{n} \mathbb{E}_{\theta_i}\left(\left\|\mathbb{E}_{\theta_i}\left[F_{t-L}^{(r)}(\theta_i)\right]-F_{t-L}^{(r)}(\theta_i)\right\|^2\right)\\
&= o(n),
\end{aligned}$$

and $\eta_G = \eta_G(\epsilon_n) = o(1)$ so that the upper bound from (3) is $2n(m+1)^2L_{\varphi}^2\eta_G = o(n)$.

Putting everything together, by Markov's inequality, we have the desired result

$$\sup_{\theta\in\Theta}\sum_{t=m+\tau+L}^{n}\left\|\mathbb{E}_{\theta}F_{t-L}^{(r)}(\theta)-F_{t-L}^{(r)}(\theta)\right\|^2 = o_p(n).$$

Step 3. We show that

$$\sup_{\theta \in \Theta} \left| I_n^3(\theta) \right| = o_p(1).$$

Observe that

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left| I_n^3(\theta) \right| \\
&\overset{(1)}{=} \sup_{\theta \in \Theta} \left| \frac{2}{n-m} \sum_{t=m+\tau+L}^{n} \hat{\mu}_{t-L,n}(\theta)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta) \right] - \mu^n_{\frac{t}{n}}(\theta) \right) \right| \\
&\overset{(2)}{\le} \sup_{\theta \in \Theta} \left| \frac{2}{n-m} \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta) \right] - \mu^n_{\frac{t}{n}}(\theta) \right) \right| \\
&+ \sup_{\theta \in \Theta} \left| \frac{2}{n-m} \sum_{t=m+\tau+L}^{n} \left(\hat{\mu}_{t-L,n}(\theta) - \mu^n_{\frac{t-L}{n}}(\theta) \right)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta) \right] - \mu^n_{\frac{t}{n}}(\theta) \right) \right|,
\end{aligned}$$

by (1) the definition of $I_n^3(\theta)$, and (2) adding and subtracting the same term, linearity in the first argument of the inner product, the triangle inequality, and the subadditivity of the supremum. We will show that both terms are $o_p(1)$.

Step 3.1. By Markov's inequality, it suffices to show that

$$\mathbb{E}\left(\sup_{\theta \in \Theta} \left| \frac{2}{n-m} \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta) \right] - \mu^n_{\frac{t}{n}}(\theta) \right) \right| \right) = o(1),$$

where $\mathbb{E}(\cdot)$ denotes expectation with respect to the law of the noise inputs. We have

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left| \frac{2}{n-m} \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta) \right] - \mu^n_{\frac{t}{n}}(\theta) \right) \right| \\
&\le \sum_{r=0}^{s} \frac{2}{n-m} \sup_{\theta \in \Theta} \left| \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left[f_t^{(r)}(\theta) - \mathbb{E}_\theta[f_t^{(r)}(\theta)] \right] \right|,
\end{aligned}$$

by linearity, the definition of the terms, the triangle inequality, upper bounding the term with the true parameter θ_0 by the supremum over $\theta \in \Theta$ of that term, and the subadditivity of the supremum. By linearity and monotonicity of expectation, it suffices to show that, for all $r = 0, 1, \ldots, s$, we have

$$\mathbb{E}\left(\sup_{\theta \in \Theta} \left| \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left[f_t^{(r)}(\theta) - \mathbb{E}_\theta[f_t^{(r)}(\theta)] \right] \right| \right) = o(n).$$

By Assumption 11, $\Theta \subset \mathbb{R}^p$ is compact, so it is totally bounded and there exists an ϵ-net, i.e. for every $\epsilon > 0$, there exist finitely many points $\theta_1, \ldots, \theta_{N(\epsilon)}$ for some $N(\epsilon) \in \mathbb{N}$, such that

$$\Theta \subset \bigcup_{i \in [N(\epsilon)]} B_\epsilon(\theta_i),$$

where $B_\epsilon(\theta_i)$ is a ball of radius ϵ. By adding and subtracting the same term, the triangle inequality, the definition of the ϵ-net for some $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, and definition of the supremum,

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left| \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left[f_t^{(r)}(\theta) - \mathbb{E}_\theta[f_t^{(r)}(\theta)] \right] \right| \\
&\overset{(1)}{\le} \max_{i \in [N(\epsilon)]} \left| \sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)] \right] \right| \\
&+ \sup_{\substack{\theta, \theta' \in \Theta \\ \|\theta - \theta'\| \le \epsilon}} \left| \sum_{t=m+\tau+L}^{n} \begin{bmatrix} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left[f_t^{(r)}(\theta) - \mathbb{E}_\theta[f_t^{(r)}(\theta)] \right] \\ - \mu^n_{\frac{t-L}{n}}(\theta')^\top \left[f_t^{(r)}(\theta') - \mathbb{E}_{\theta'}[f_t^{(r)}(\theta')] \right] \end{bmatrix} \right|.
\end{aligned}$$

Taking the expectation $\mathbb{E}(\cdot)$ with respect to the law of the noise inputs, we have

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta)^\top\left[f_t^{(r)}(\theta)-\mathbb{E}_\theta[f_t^{(r)}(\theta)]\right]\right|\right)\\
&\overset{(1)}{\le}\mathbb{E}\left(\max_{i\in[N(\epsilon)]}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta_i)^\top\left[f_t^{(r)}(\theta_i)-\mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right|\right)\\
&+\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\left|\sum_{t=m+\tau+L}^{n}\begin{bmatrix}\mu^n_{\frac{t-L}{n}}(\theta)^\top\left[f_t^{(r)}(\theta)-\mathbb{E}_\theta[f_t^{(r)}(\theta)]\right]\\-\mu^n_{\frac{t-L}{n}}(\theta')^\top\left[f_t^{(r)}(\theta')-\mathbb{E}_{\theta'}[f_t^{(r)}(\theta')]\right]\end{bmatrix}\right|\right]\\
&\overset{(2)}{\le}\sum_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta_i)^\top\left[f_t^{(r)}(\theta_i)-\mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right|\\
&+\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\left|\sum_{t=m+\tau+L}^{n}\begin{bmatrix}\mu^n_{\frac{t-L}{n}}(\theta)^\top\left[f_t^{(r)}(\theta)-\mathbb{E}_\theta[f_t^{(r)}(\theta)]\right]\\-\mu^n_{\frac{t-L}{n}}(\theta')^\top\left[f_t^{(r)}(\theta')-\mathbb{E}_{\theta'}[f_t^{(r)}(\theta')]\right]\end{bmatrix}\right|\right]\\
&\overset{(3)}{\le}\sum_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta_i)^\top\left[f_t^{(r)}(\theta_i)-\mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right|\\
&+\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta')\right\|\left\|f_t^{(r)}(\theta)-\mathbb{E}_\theta[f_t^{(r)}(\theta)]\right\|\right]\\
&+\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta')\right\|\left\|(f_t^{(r)}(\theta)-f_t^{(r)}(\theta'))-(\mathbb{E}_\theta[f_t^{(r)}(\theta)]-\mathbb{E}_{\theta'}[f_t^{(r)}(\theta')])\right\|\right]\\
&\overset{(4)}{\le}\sum_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta_i)^\top\left[f_t^{(r)}(\theta_i)-\mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right|\\
&+2\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|\mu^n_{\frac{t-L}{n}}(\theta)-\mu^n_{\frac{t-L}{n}}(\theta')\right\|\\
&+2\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|f_t^{(r)}(\theta)-f_t^{(r)}(\theta')\right\|\right]\\
&\overset{(5)}{\le}\sum_{i\in[N(\epsilon)]}\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n}\mu^n_{\frac{t-L}{n}}(\theta_i)^\top\left[f_t^{(r)}(\theta_i)-\mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right|\\
&+2(s+3)\mathbb{E}\left[\sup_{\substack{\theta,\theta'\in\Theta\\\|\theta-\theta'\|\le\epsilon}}\sum_{t=m+\tau+L}^{n}\left\|f_t^{(r)}(\theta)-f_t^{(r)}(\theta')\right\|\right],
\end{aligned}
$$

by (1) the previous ϵ-net inequality, monotonicity of expectation, and linearity of expectation, (2) upper bounding the maximum by the sum and linearity of expectation, (3) adding and subtracting the cross-term with θ and θ', the triangle inequality, the Cauchy-Schwarz inequality, and the subadditivity of the supremum, (4) upper bounding because the random features are bounded between -1 and 1 so the expectations must be as well, and for the last term: the triangle inequality, linearity, subadditivity of the supremum, linearity of expectation, Jensen's inequality, and monotonicity of expectation, and (5) linearity of expectation, Jensen's inequality, the triangle inequality, the Minkowski inequality, and subadditivity of the supremum. Next, we show both terms are $o(n)$.

Step 3.1.1. We begin with the first term. For each $r=0,1,\dots,s$, and $\theta\in\Theta$, denote

$$
Y_t^{(r)}(\theta)=\mu^n_{\frac{t-L}{n}}(\theta)^\top\left[f_t^{(r)}(\theta)-\mathbb{E}_\theta[f_t^{(r)}(\theta)]\right],\quad t=m+\tau+L,\dots,n,
$$

which admits the representation

$$Y_t^{(r)}(\theta) = H_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta),$$

where the mapping $H_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta)$ is defined in terms of the original causal representations from Assumption 8,

$$H_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta) = \mu_{\frac{t-L}{n}}^{n}(\theta)^\top \left[\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta)\right]^\top\right) - \mathbb{E}_\theta[f_t^{(r)}(\theta)]\right].$$

Analogous to Assumption 15, denote the version of $Y_t^{(r)}(\theta)$ with the j-th noise input in the past replaced with its iid copy by

$$\begin{aligned}\tilde{Y}_{t,j}^{(r)}(\theta) &= H_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}^{(r)}, \theta)\\ &= \mu_{\frac{t-L}{n}}^{n}(\theta)^\top \left[\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\tilde{\varepsilon}_{t-m,j-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}^{(r)}, \theta)\right]^\top\right) - \mathbb{E}_\theta[f_t^{(r)}(\theta)]\right],\end{aligned}$$

where for any $i \in \mathbb{N}$ we define $\tilde{\varepsilon}_{t-i,j-i}^{(r)}$ as $\varepsilon_{t-i}^{(r)}$ if $j - i \leq 0$. Observe that the time series $Y_t^{(r)}(\theta)$ satisfies a decaying physical dependence measure condition

$$\begin{aligned}&\left\|Y_t^{(r)}(\theta) - \tilde{Y}_{t,j}^{(r)}(\theta)\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(1)}{=} \left\|H_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta) - H_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}^{(r)}, \theta)\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(2)}{=} \left\|\mu_{\frac{t-L}{n}}^{n}(\theta)^\top \begin{bmatrix}\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta)\right]^\top\right)\\ -\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\tilde{\varepsilon}_{t-m,j-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}^{(r)}, \theta)\right]^\top\right)\end{bmatrix}\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(3)}{\leq} \left\|\mu_{\frac{t-L}{n}}^{n}(\theta)\right\| \left\|\begin{matrix}\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\varepsilon_{t-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\varepsilon_t^{(r)}, \theta)\right]^\top\right)\\ -\varphi\left(\left[G_{(t-m)/n}^{(n,r)}(\tilde{\varepsilon}_{t-m,j-m}^{(r)}, \theta), \ldots, G_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}^{(r)}, \theta)\right]^\top\right)\end{matrix}\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(4)}{\leq} L_\varphi \sum_{i=0}^{m} \left\|G_{(t-i)/n}^{(n,r)}(\varepsilon_{t-i}^{(r)}, \theta) - G_{(t-i)/n}^{(n,r)}(\tilde{\varepsilon}_{t-i,j-i}^{(r)}, \theta)\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(5)}{=} L_\varphi \sum_{i=0}^{\min(m,j-1)} \left\|G_{(t-i)/n}^{(n,r)}(\varepsilon_{t-i}^{(r)}, \theta) - G_{(t-i)/n}^{(n,r)}(\tilde{\varepsilon}_{t-i,j-i}^{(r)}, \theta)\right\|_{\mathcal{L}^q(\theta)}\\
&\overset{(6)}{\leq} L_\varphi \sum_{i=0}^{\min(m,j-1)} \Lambda \rho^{j-i}\\
&\overset{(7)}{\leq} L_\varphi (m+1) \Lambda \rho^{j-\min(m,j-1)}\\
&\overset{(8)}{\leq} \Lambda^H (\rho^H)^j,\end{aligned}$$

by (1) definition, (2) cancellation of terms and the linearity of the inner product, (3) the Cauchy-Schwarz inequality and linearity of expectation, (4) because the random Fourier features are bounded between -1 and 1 so the expectations must be as well, and because the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, and the Minkowski inequality, (5) because the distances are zero for $j - i \leq 0$, so it suffices to consider the terms where $j - i > 0 \iff i \leq j - 1$, (6) using the bound on the physical dependence measure from Assumption 15, (7) upper bounding by $m + 1$ times the largest term which is the "earliest" term because $\rho \in (0, 1)$, and (8) upper bounding with an expression that satisfies exponential decay condition analogous to Assumption 15, where $\Lambda^H = L_\varphi(m + 1)\Lambda > 0$ and $\rho^H = \rho \in (0, 1)$. Lemma 7 is therefore applicable to the time series $Y_t^{(r)}(\theta)$.

Let us return to the term of interest. For all $i \in N(\epsilon)$, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right| \\
&\overset{(1)}{=} \mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} Y_t^{(r)}(\theta_i)\right| \\
&\overset{(2)}{\leq} Cn^{\frac{1}{2}}\Lambda^H \sum_{j=1}^{\infty}(\rho^H)^j \\
&\overset{(3)}{=} Cn^{\frac{1}{2}}\Lambda^H K^H,
\end{aligned}
$$

by (1) definition, (2) the second inequality from Lemma 7, noting that $\mathbb{E}_\theta[Y_t^{(r)}(\theta)] = 0$ for all t, r, θ, by construction, applying the inequality on the physical dependence measure derived above, and further upper bounding with infinitely many terms in the summation, (3) noting that the geometric series converges and the constant $K^H = \frac{1}{1-\rho^H} - 1 > 0$ since $\rho^H = \rho \in (0,1)$ by Assumption 15. Hence, for each fixed $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, for all $i \in N(\epsilon)$, we have

$$
\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right| = O(n^{\frac{1}{2}}).
$$

Step 3.1.2. We now consider the second term. Observe that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{\theta,\theta' \in \Theta \\ \|\theta-\theta'\| \leq \epsilon}} \sum_{t=m+\tau+L}^{n} \left\|f_t^{(r)}(\theta) - f_t^{(r)}(\theta')\right\|\right] \\
&\overset{(1)}{\leq} L_\varphi \sum_{t=m+\tau+L}^{n} \sum_{j=0}^{m} \mathbb{E}\left(\sup_{\substack{\theta,\theta' \in \Theta \\ \|\theta-\theta'\| \leq \epsilon}} \left\|G^{(n,r)}_{(t-j)/n}(\varepsilon^{(r)}_{t-j}, \theta) - G^{(n,r)}_{(t-j)/n}(\varepsilon^{(r)}_{t-j}, \theta')\right\|\right) \\
&\overset{(2)}{\leq} L_\varphi n(m+1)\eta_G,
\end{aligned}
$$

because (1) the random Fourier features φ are L_φ-Lipschitz with Lipschitz constant given by Eq. (26), the triangle inequality, subadditivity of the supremum, and linearity of expectation, and (2) by upper bounding by n times the supremum over $u \in [0,1]$ and applying the stochastic equicontinuity-type condition from Assumption 14. By Assumption 14, $\eta_G \to 0$ as $\epsilon \to 0$. However, taking $\epsilon \to 0$ makes $N(\epsilon) \to \infty$, so we must specify the rates.

Previously, we showed that, for each fixed $\epsilon > 0$ and $N(\epsilon) \in \mathbb{N}$, for all $i \in N(\epsilon)$, we have

$$
\mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right| = O(n^{\frac{1}{2}}).
$$

Therefore, for any sequence $(\epsilon_n)_{n\in\mathbb{N}}$, $\epsilon_n \to 0$ as $n \to \infty$, such that $N(\epsilon_n) = o(n^{\frac{1}{2}})$, we have

$$
\begin{aligned}
&\sum_{i\in[N(\epsilon)]} \mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right| \\
&\leq N(\epsilon) \max_{i\in[N(\epsilon)]} \mathbb{E}_{\theta_i}\left|\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta_i)^\top \left[f_t^{(r)}(\theta_i) - \mathbb{E}_{\theta_i}[f_t^{(r)}(\theta_i)]\right]\right| \\
&= o(n),
\end{aligned}
$$

and $\eta_G = \eta_G(\epsilon_n) = o(1)$ so that the upper bound from (2) is $L_\varphi n(m+1)\eta_G = o(n)$.

Putting everything together, by Markov's inequality, we have the desired result

$$
\sup_{\theta\in\Theta}\left|\frac{2}{n-m}\sum_{t=m+\tau+L}^{n} \mu^n_{\frac{t-L}{n}}(\theta)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta)\right] - \mu^n_{\frac{t}{n}}(\theta)\right)\right| = o_p(1).
$$

Step 3.2. Next, we show that

$$
\sup_{\theta\in\Theta}\left|\frac{2}{n-m}\sum_{t=m+\tau+L}^{n} \left(\hat{\mu}_{t-L,n}(\theta) - \mu^n_{\frac{t-L}{n}}(\theta)\right)^\top \left(\left[f_t^{\text{obs}} - \bar{f}_t^{\text{sim}}(\theta)\right] - \mu^n_{\frac{t}{n}}(\theta)\right)\right| = o_p(1).
$$

Observe that

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left(\hat{\mu}_{t-L,n}(\theta)-\mu^{n}_{\frac{t-L}{n}}(\theta)\right)^{\top}\left(\left[f^{\text{obs}}_t-\bar{f}^{\text{sim}}_t(\theta)\right]-\mu^{n}_{\frac{t}{n}}(\theta)\right)\right| \\
&\overset{(1)}{\le}\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\hat{\mu}_{t-L,n}(\theta)-\mu^{n}_{\frac{t-L}{n}}(\theta)\right\|\left\|\left(\left[f^{\text{obs}}_t-\bar{f}^{\text{sim}}_t(\theta)\right]-\mu^{n}_{\frac{t}{n}}(\theta)\right)\right\| \\
&\overset{(2)}{\le}\left(\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\hat{\mu}_{t-L,n}(\theta)-\mu^{n}_{\frac{t-L}{n}}(\theta)\right\|^{2}\right)^{\frac{1}{2}} \\
&\left(\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\left(\left[f^{\text{obs}}_t-\bar{f}^{\text{sim}}_t(\theta)\right]-\mu^{n}_{\frac{t}{n}}(\theta)\right)\right\|^{2}\right)^{\frac{1}{2}} \\
&\overset{(3)}{\le}C\left(\sup_{\theta\in\Theta}\frac{2}{n-m}\sum_{t=m+\tau+L}^{n}\left\|\hat{\mu}_{t-L,n}(\theta)-\mu^{n}_{\frac{t-L}{n}}(\theta)\right\|^{2}\right)^{\frac{1}{2}} \\
&\overset{(4)}{=}o_p(1),
\end{aligned}
$$

by (1) the triangle inequality and the Cauchy-Schwarz inequality, (2) the Cauchy-Schwarz inequality, (3) the triangle inequality and the fact that the random features are bounded between -1 and 1, so we can upper bound this average by some constant $C>0$, and (4) because this term was previously shown to be $o_p(1)$ in Step 2.2 of the proof Eq. (25). □

2.7. Auxiliary results. To begin, we derive the Lipschitz constant for the random Fourier features.

Lemma 1. *The function $\varphi=(\varphi_1,\dots,\varphi_k)$ from* Eq. (2) *is Lipschitz with Lipschitz constant*

$$
L_{\varphi}=\left(\sum_{i=1}^{k}\sum_{j=1}^{m+1}\|\Omega_{i,j}\|^{2}\right)^{1/2}. \qquad \textbf{[26]}
$$

Proof: For each coordinate $i\in[k]$ of $\varphi=(\varphi_1,\dots,\varphi_k)$, we have, for $x,y\in\mathbb{R}^{(m+1)\times d}$,

$$
\begin{aligned}
&|\varphi_i(x)-\varphi_i(y)| \\
&\overset{(1)}{=}\left|\cos\left(\sum_{j=1}^{m+1}\Omega_{i,j}\cdot x_j+\alpha_i\right)-\cos\left(\sum_{j=1}^{m+1}\Omega_{i,j}\cdot y_j+\alpha_i\right)\right| \\
&\overset{(2)}{\le}\left|\sum_{j=1}^{m+1}\Omega_{i,j}\cdot(x_j-y_j)\right| \\
&\overset{(3)}{\le}\sum_{j=1}^{m+1}\|\Omega_{i,j}\|\,\|x_j-y_j\| \\
&\overset{(4)}{\le}\left(\sum_{j=1}^{m+1}\|\Omega_{i,j}\|^{2}\right)^{1/2}\left(\sum_{j=1}^{m+1}\|x_j-y_j\|^{2}\right)^{1/2} \\
&\overset{(5)}{\le}\left(\sum_{j=1}^{m+1}\|\Omega_{i,j}\|^{2}\right)^{1/2}\|\text{Vec}(x)-\text{Vec}(y)\|,
\end{aligned}
$$

because (1) the definition of the random Fourier features φ, (2) cosine is 1-Lipschitz, canceling the phase α_i, and linearity of sums and inner products, (3) the triangle inequality and the Cauchy-Schwarz inequality applied to each inner product, (4) the Cauchy-Schwarz inequality applied to the sum of products, and (5) definition of the Euclidean norm, which yields

$$
|\varphi_i(x)-\varphi_i(y)|\le\left(\sum_{j=1}^{m+1}\|\Omega_{i,j}\|^{2}\right)^{1/2}\|\text{Vec}(x)-\text{Vec}(y)\|,
$$

so squaring both sides, summing over $i = 1, \dots, k$, and taking square roots, yields

$$\|\varphi(x) - \varphi(y)\| \le \left(\sum_{i=1}^{k} \sum_{j=1}^{m+1} \|\Omega_{i,j}\|^2 \right)^{1/2} \|\mathrm{Vec}(x) - \mathrm{Vec}(y)\| .$$

□

For $t \in \mathbb{N}$, $i \in \mathbb{N}$, $j \in \mathbb{N}_0$, $r = 0, 1, \dots, s$, define

$$Z_t^{(i,r)}(\theta) = \left[G_{t-m}^{(i,r)}(\varepsilon_{-m}^{(r)}, \theta), \dots, G_t^{(i,r)}(\varepsilon_0^{(r)}, \theta) \right]^\top ,$$
$$\tilde{Z}_{t,j}^{(i,r)}(\theta) = \left[G_{t-m}^{(i,r)}(\tilde{\varepsilon}_{-m,j}^{(r)}, \theta), \dots, G_t^{(i,r)}(\tilde{\varepsilon}_{0,j}^{(r)}, \theta) \right]^\top ,$$

where the definition of $\tilde{\varepsilon}_{\ell,j}$, $\ell \in \mathbb{Z}$, is from Eq. (11).

The following result establishes that the time series of random features satisfies a physical dependence measure condition, analogous to Assumption 7.

Lemma 2. *There exist $\Psi^\varphi > 0$, $\beta^\varphi > 2$ such that, for some $q^\varphi > 2$, all $j \in \mathbb{N}_0$, all $i \in \mathbb{N}$, and all $r = 0, 1, \dots, s$, we have*

$$\sup_{\theta \in \Theta} \sup_{t \in \mathbb{N}} \left\| \varphi\left(Z_t^{(i,r)}(\theta)\right) - \varphi\left(\tilde{Z}_{t,j}^{(i,r)}(\theta)\right) \right\|_{\mathcal{L}^{q^\varphi}(\theta)} \le \Psi^\varphi (j \vee 1)^{-\beta^\varphi},$$
$$\sup_{\theta \in \Theta} \sup_{t \in \mathbb{N}} \left\| \varphi\left(Z_t^{(i,r)}(\theta)\right) \right\|_{\mathcal{L}^{q^\varphi}(\theta)} \le \Psi^\varphi .$$

Proof: The result follows by the fact that φ is Lipschitz with Lipschitz constant L_φ from Eq. (26), the triangle inequality and Minkowski inequality, and applying the bounds from Assumption 7. The same line of reasoning is used in the proof of Theorem 2, so the details are omitted.

□

The following result presents Theorem 3.2 from (11) using our notation.

Lemma 3. (Law of Large Numbers (11)). *For $r = 0, 1, \dots, s$ and $\theta \in \Theta$, suppose a time series $X_t^{(r)}(\theta)$, $t = 1, \dots, n$, is generated as*

$$X_t^{(r)}(\theta) = G_t^{(n,r)}(\varepsilon_t, \theta).$$

Suppose for all $i \in \mathbb{N}$ and $r = 0, 1, \dots, s$, there exists a $q > 2$, such that

$$\sup_{\theta \in \Theta} \sup_{t \in \mathbb{N}} \left\| G_t^{(i,r)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} < \infty .$$

For some $2 \le h \le q < \infty$, define the physical dependence measure at time t and sample size n as

$$\eta_{t,j,q,h}^{(n,r)}(\theta) = \mathbb{E}_\theta \left(\left\| G_t^{(n,r)}(\varepsilon_t, \theta) - G_t^{(n,r)}(\tilde{\varepsilon}_{t,j}, \theta) \right\|_h^q \right)^{1/q} .$$

There exists a constant $C = C(q, h, \theta)$, such that for all $r = 0, 1, \dots, s$, $\theta \in \Theta$, and $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\left(\mathbb{E}_\theta \left[\max_{\ell \le n} \left\| \sum_{t=1}^{\ell} \left(X_t^{(r)}(\theta) - \mathbb{E}_\theta \left[X_t^{(r)}(\theta) \right] \right) \right\|_h^q \right] \right)^{\frac{1}{q}} \\
&= \left(\mathbb{E}_\theta \left[\max_{\ell \le n} \left\| \sum_{t=1}^{\ell} \left(G_t^{(n,r)}(\varepsilon_t, \theta) - \mathbb{E}_\theta \left[G_t^{(n,r)}(\varepsilon_t, \theta) \right] \right) \right\|_h^q \right] \right)^{\frac{1}{q}} \\
&\le C n^{\frac{1}{2} - \frac{1}{q}} \sum_{j=1}^{\infty} \left(\sum_{t=1}^{n} (\eta_{t,j,q,h}^{(n,r)}(\theta))^q \right)^{\frac{1}{q}} \\
&\le C n^{\frac{1}{2}} \sum_{j=1}^{\infty} \max_{t \le n} \eta_{t,j,q,h}^{(n,r)}(\theta),
\end{aligned}$$

and in the special case of $h = 2$, the inequality may be improved to

$$\left(\mathbb{E}_\theta\left[\max_{\ell\le n}\left\|\sum_{t=1}^{\ell}\left(X_t^{(r)}(\theta)-\mathbb{E}_\theta\left[X_t^{(r)}(\theta)\right]\right)\right\|_2^q\right]\right)^{\frac{1}{q}}$$
$$=\left(\mathbb{E}_\theta\left[\max_{\ell\le n}\left\|\sum_{t=1}^{\ell}\left(G_t^{(n,r)}(\varepsilon_t,\theta)-\mathbb{E}_\theta\left[G_t^{(n,r)}(\varepsilon_t,\theta)\right]\right)\right\|_2^q\right]\right)^{\frac{1}{q}}$$
$$\le C\sum_{j=1}^{\infty}(j\wedge n)^{\frac{1}{2}-\frac{1}{q}}\left(\sum_{t=1}^{n}(\eta_{t,j,q,2}^{(n,r)}(\theta))^q\right)^{\frac{1}{q}}+C\sum_{j=1}^{n}\left(\sum_{t=1}^{n}(\eta_{t,j,2,2}^{(n,r)}(\theta))^2\right)^{\frac{1}{2}}.$$

Proof: See the proof of Theorem 3.2 in (11). □

For $u\in[0,1]$, $i\in\mathbb{N}$, $j\in\mathbb{N}_0$, $r=0,1,\ldots,s$, define

$$Z_u^{(i,r)}(\theta)=\left[G_u^{(i,r)}(\varepsilon_{-m}^{(r)},\theta),\ldots,G_u^{(i,r)}(\varepsilon_0^{(r)},\theta)\right]^\top,$$
$$\tilde{Z}_{u,j}^{(i,r)}(\theta)=\left[G_u^{(i,r)}(\tilde{\varepsilon}_{-m,j}^{(r)},\theta),\ldots,G_u^{(i,r)}(\tilde{\varepsilon}_{0,j}^{(r)},\theta)\right]^\top,$$

where the definition of $\tilde{\varepsilon}_{\ell,j}$, $\ell\in\mathbb{Z}$, is from Eq. (11).

The next result shows that the time series of random features satisfies a physical dependence measure condition, analogous to Assumption 15.

Lemma 4. *There exist constants $\rho^\varphi\in(0,1)$, $\Lambda^\varphi>0$, such that, for some $q^\varphi>2$, all $j\in\mathbb{N}_0$, all $i\in\mathbb{N}$, and all $r=0,1,\ldots,s$, we have*

$$\sup_{\theta\in\Theta}\sup_{u\in[0,1]}\left\|\varphi\left(Z_u^{(i,r)}(\theta)\right)-\varphi\left(\tilde{Z}_{u,j}^{(i,r)}(\theta)\right)\right\|_{\mathcal{L}^{q^\varphi}(\theta)}\le\Lambda^\varphi(\rho^\varphi)^j,$$
$$\sup_{\theta\in\Theta}\sup_{u\in[0,1]}\left\|\varphi\left(Z_u^{(i,r)}(\theta)\right)\right\|_{\mathcal{L}^{q^\varphi}(\theta)}\le\Lambda^\varphi.$$

Proof: The result follows by the fact that φ is Lipschitz with Lipschitz constant L_φ from Eq. (26), the triangle inequality and Minkowski inequality, and applying the bounds from Assumption 15. The same line of reasoning is used in the proof of Theorem 5, so the details are omitted. □

The following result establishes that the time series of random features satisfies a nonstationarity condition, analogous to Assumption 16.

Lemma 5. *For some $q^\varphi>2$, some $\kappa^\varphi\in[1,4)$, all $i\in\mathbb{N}$, and all $r=0,1,\ldots,s$, the constant $\Lambda^\varphi>0$ from Lemma 4 also satisfies*

$$\sup_{\theta\in\Theta}\sup_{u\in[0,1]}\left\|\varphi\left(Z_u^{(i,r)}(\theta)\right)\right\|_{\mathcal{L}^{q^\varphi}(\theta)}+\sup_{\theta\in\Theta}\left\|\left(\varphi\left(Z_u^{(i,r)}(\theta)\right)\right)_u\right\|_{\kappa^\varphi\text{-var},\mathcal{L}^{q^\varphi}(\theta)}\le\Lambda^\varphi.$$

Proof: The result follows by the fact that φ is Lipschitz with Lipschitz constant L_φ from Eq. (26), the triangle inequality and Minkowski inequality, and applying the bounds from Assumption 16. The same line of reasoning is used in the proof of Theorem 5, so the details are omitted. □

For all times $t=m+1,\ldots,n$, simulation indices $r=0,1,\ldots,s$, and parameter values $\theta\in\Theta$, recall the r-th random feature $f_t^{(r)}(\theta)$ at time t based on θ as in Eq. (3). The time series of random features satisfies a decaying autocovariance condition, analogous to Proposition 1 from (12). For the following result, denote the trace norm of a matrix $A\in\mathbb{R}^{k\times k}$ by $\|A\|_{\mathrm{tr}}=\mathrm{tr}((A^\top A)^{\frac{1}{2}})$.

Lemma 6. *There exist constants $C^\varphi\equiv C^\varphi(\rho^\varphi,\Lambda^\varphi)>0$, $K^\varphi\equiv K^\varphi(\rho^\varphi,\Lambda^\varphi)>1$ depending on the constants $\rho^\varphi\in(0,1)$, $\Lambda^\varphi>0$ from Lemmas 4 and 5 such that, for all $t=m+1,\ldots,n$ and $r=0,\ldots,s$, we have*

$$\sup_{\theta\in\Theta}\left\|\mathrm{Cov}\left(f_{t_2}^{(r)}(\theta),f_{t_1}^{(r)}(\theta)\right)\right\|_{\mathrm{tr}}\le C^\varphi(|t_2-t_1|+1)^{-K^\varphi}.$$

Proof: Lemma 4 states that, uniformly over the parameter values, the physical dependence measure of the time series of random features decays exponentially. In particular, this implies a (slower) polynomial decay of the physical dependence measure of the random features as in Assumption A.2 of (12). That is, there exist constants $A^{\varphi} \equiv A^{\varphi}(\rho^{\varphi}, \Lambda^{\varphi}) > 0$, $K^{\varphi} \equiv K^{\varphi}(\rho^{\varphi}, \Lambda^{\varphi}) > 1$ such that, for some $q^{\varphi} > 2$, all $j \in \mathbb{N}_0$, all $i \in \mathbb{N}$, and all $r = 0, 1, \ldots, s$, we have

$$\sup_{\theta \in \Theta} \sup_{u \in [0,1]} \left\| \varphi\left(Z_u^{(i,r)}(\theta)\right) - \varphi\left(\tilde{Z}_{u,j}^{(i,r)}(\theta)\right) \right\|_{\mathcal{L}^{q^{\varphi}}(\theta)} \leq A^{\varphi}(j+1)^{-K^{\varphi}}.$$

Proposition 1 in (12) proves that a polynomial decay of the physical dependence measure is sufficient to imply a decay of the trace norm of the autocovariance matrix of a time series. The same arguments used to prove Proposition 1 in (12) can be applied to the time series of random features for each parameter value, i.e. for all $t = m+1, \ldots, n$, $r = 0, \ldots, s$, and $\theta \in \Theta$ we have

$$\left\| \mathrm{Cov}\left(f_{t_2}^{(r)}(\theta), f_{t_1}^{(r)}(\theta)\right) \right\|_{\mathrm{tr}} \leq C^{\varphi}(|t_2 - t_1| + 1)^{-K^{\varphi}}.$$

Since the suprema over all $\theta \in \Theta$ of the upper bounds are always finite, the Θ-uniform inequality from the Lemma hold by basic properties of the supremum. □

The following result is Theorem 3.2 from (11) using our notation, which is possible because the triangle array framework of (11) nests the locally stationary time series setting; see Examples 2 and 3 therein.

Lemma 7. (Law of Large Numbers (11)). *For $r = 0, 1, \ldots, s$ and $\theta \in \Theta$, suppose a time series $X_t^{(r)}(\theta)$, $t = 1, \ldots, n$, is generated as*

$$X_t^{(r)}(\theta) = G_{t/n}^{(n,r)}(\varepsilon_t, \theta).$$

Suppose for all $i \in \mathbb{N}$ and $r = 0, 1, \ldots, s$, there exists a $q > 2$, such that

$$\sup_{\theta \in \Theta} \sup_{u \in [0,1]} \left\| G_u^{(i,r)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} < \infty.$$

For some $2 \leq h \leq q < \infty$, define the physical dependence measure at time t and sample size n as

$$\eta_{t,j,q,h}^{(n,r)}(\theta) = \mathbb{E}_{\theta}\left(\left\| G_{t/n}^{(n,r)}(\varepsilon_t, \theta) - G_{t/n}^{(n,r)}(\tilde{\varepsilon}_{t,j}, \theta) \right\|_h^q\right)^{1/q}.$$

There exists a constant $C = C(q, h, \theta)$, such that for all $r = 0, 1, \ldots, s$, $\theta \in \Theta$, and $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\left(\mathbb{E}_{\theta}\left[\max_{\ell \leq n} \left\| \sum_{t=1}^{\ell} \left(X_t^{(r)}(\theta) - \mathbb{E}_{\theta}\left[X_t^{(r)}(\theta)\right]\right) \right\|_h^q\right]\right)^{\frac{1}{q}} \\
&= \left(\mathbb{E}_{\theta}\left[\max_{\ell \leq n} \left\| \sum_{t=1}^{\ell} \left(G_{t/n}^{(n,r)}(\varepsilon_t, \theta) - \mathbb{E}_{\theta}\left[G_{t/n}^{(n,r)}(\varepsilon_t, \theta)\right]\right) \right\|_h^q\right]\right)^{\frac{1}{q}} \\
&\leq C n^{\frac{1}{2} - \frac{1}{q}} \sum_{j=1}^{\infty} \left(\sum_{t=1}^{n} (\eta_{t,j,q,h}^{(n,r)}(\theta))^q\right)^{\frac{1}{q}} \\
&\leq C n^{\frac{1}{2}} \sum_{j=1}^{\infty} \max_{t \leq n} \eta_{t,j,q,h}^{(n,r)}(\theta),
\end{aligned}$$

and in the special case of $h = 2$, the inequality may be improved to

$$\begin{aligned}
&\left(\mathbb{E}_{\theta}\left[\max_{\ell \leq n} \left\| \sum_{t=1}^{\ell} \left(X_t^{(r)}(\theta) - \mathbb{E}_{\theta}\left[X_t^{(r)}(\theta)\right]\right) \right\|_2^q\right]\right)^{\frac{1}{q}} \\
&= \left(\mathbb{E}_{\theta}\left[\max_{\ell \leq n} \left\| \sum_{t=1}^{\ell} \left(G_{t/n}^{(n,r)}(\varepsilon_t, \theta) - \mathbb{E}_{\theta}\left[G_{t/n}^{(n,r)}(\varepsilon_t, \theta)\right]\right) \right\|_2^q\right]\right)^{\frac{1}{q}} \\
&\leq C \sum_{j=1}^{\infty} (j \wedge n)^{\frac{1}{2} - \frac{1}{q}} \left(\sum_{t=1}^{n} (\eta_{t,j,q,2}^{(n,r)}(\theta))^q\right)^{\frac{1}{q}} + C \sum_{j=1}^{n} \left(\sum_{t=1}^{n} (\eta_{t,j,2,2}^{(n,r)}(\theta))^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Proof: See the proof of Theorem 3.2 in (11). □

3. Additional experiments

3.1. Time-average estimator.

Gaussian model. For $t = 1, \ldots, n$, we observe

$$X_t = \mu + \varepsilon_t,$$

where each $\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1)$. The unknown parameter is μ, so we use $2p + 1 = 3$ random features. We aim to estimate

$$\mu_0 = 0.5.$$

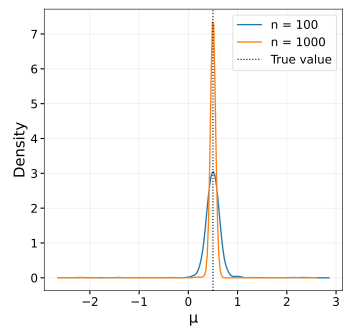

Fig. 5. Gaussian model. Density of parameter estimates using $2p + 1 = 3$ random features.

Moving average process. For $t = 1, \ldots, n$, we observe

$$X_t = \mu + \psi\varepsilon_{t-1} + \varepsilon_t,$$

where the noise inputs ε_t are sampled iid from the intractable g-and-k distribution. For more information about the g-and-k distribution, see (13, 14). We set the location component to zero, the kurtosis component to 0.1, and the c component to 0.8. The remaining components σ and g control the scale and skewness, respectively. We aim to estimate

$$\psi_0 = 0.5, \quad \mu_0 = -3, \quad \sigma_0 = 0.2, \quad g_0 = 0.1.$$

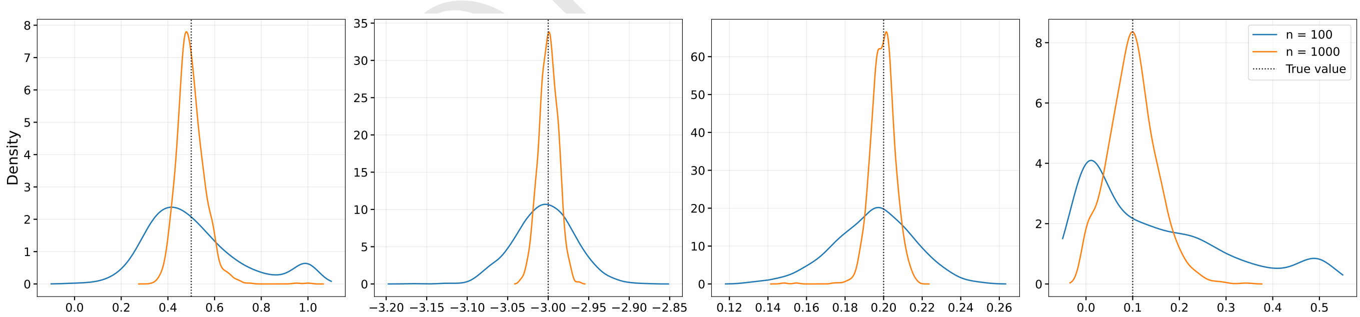

Fig. 6. Moving average process. Density of parameter estimates using $2p + 1 = 9$ random features.

Autoregressive process. For $t = 1, \ldots, n$, we observe

$$X_t = X_{t-1} + \psi(\mu - X_{t-1}) + \varepsilon_t,$$

where the noise inputs ε_t are sampled iid from the intractable g-and-k distribution and the initial value $X_0 = 3$ is known, so the model may be viewed as a discretized version of a Lévy-driven Ornstein–Uhlenbeck process. We set the location component to zero, the kurtosis component to 0.2, and c component to 0.8. The remaining components σ and g control the scale and skewness. We aim to estimate

$$\psi_0 = 0.3, \quad \mu_0 = 2, \quad \sigma_0 = 0.3, \quad g_0 = 0.2.$$

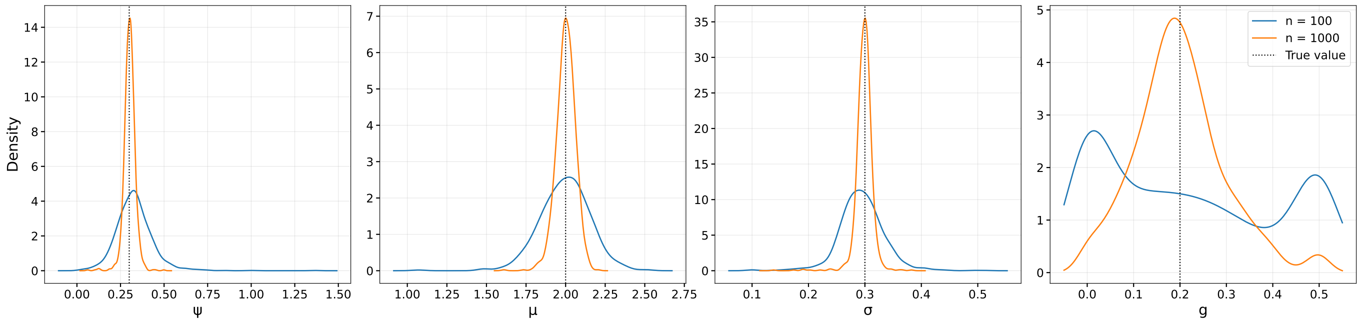

Fig. 7. Autoregressive process. Density of parameter estimates using $2p+1=9$ random features.

State-space model. For times $t = 1, \ldots, n$, define the latent state by

$$Z_t = Z_{t-1} + \psi(\mu - Z_{t-1}) + \sigma\varepsilon_t,$$

where each $\varepsilon_t \overset{\text{iid}}{\sim} N(0,1)$. For dimensions $j = 1, \ldots, d \equiv 25$, we observe

$$X_t^{(j)} = Z_t + \lambda\xi_t^{(j)},$$

where each $\xi_t^{(j)} \overset{\text{iid}}{\sim} N(0,1)$. We aim to estimate

$$\psi_0 = 0.25, \quad \sigma_0 = 0.08, \quad \mu_0 = 2.5, \quad \lambda_0 = 0.05.$$

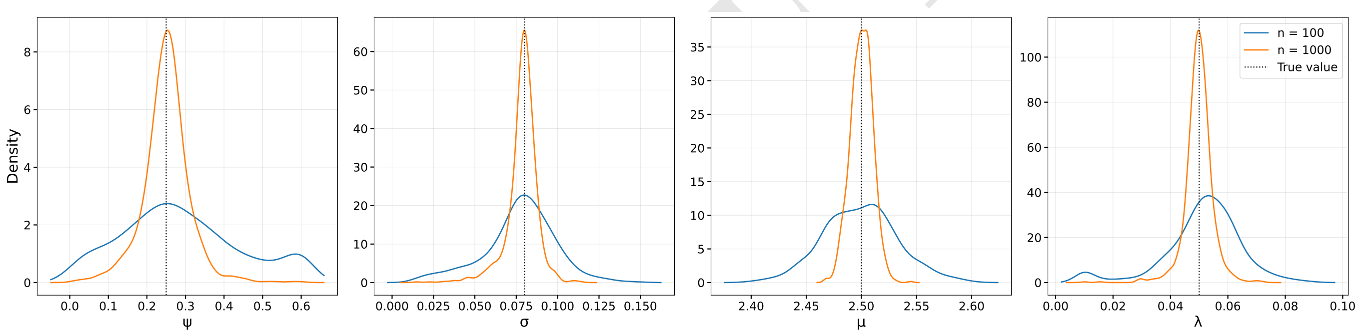

Fig. 8. State-space model. Density of parameter estimates using $2p+1=9$ random features.

3.2. Rolling-window estimator.

Structural time series model. We consider a two-dimensional structural time series model with cycles, trends, and an abrupt change-point. For $u \in [0,1]$, define

$$Z^{(1)}(u) = \mu^{(1)}u + \alpha\mathbf{1}_{u\geq\tau} + \cos(2\pi\beta^{(1)}u), \quad Z^{(2)}(u) = \mu^{(2)}u - \alpha\mathbf{1}_{u\geq\tau} + \sin(2\pi\beta^{(2)}u).$$

For $t = 1, \ldots, n$, we observe

$$X_t^{(1)} = Z^{(1)}(t/n) + \sigma^{(1)}\varepsilon_t^{(1)}, \quad X_t^{(2)} = Z^{(2)}(t/n) + \sigma^{(2)}\varepsilon_t^{(2)},$$

where each $\varepsilon_t^{(1)}, \varepsilon_t^{(2)} \overset{\text{iid}}{\sim} N(0,1)$ and each $\sigma^{(1)}, \sigma^{(2)} > 0$ controls the level of the noise. We aim to estimate

$$\alpha_0 = 1.8, \quad \tau_0 = 0.28, \quad \mu_0^{(1)} = -1.5, \quad \mu_0^{(2)} = 0.4,$$
$$\beta_0^{(1)} = 2.8, \quad \beta_0^{(2)} = 2.8, \quad \sigma_0^{(1)} = 0.9, \quad \sigma_0^{(2)} = 0.9.$$

4. Related work

We provide a selective overview of relevant strands of literature.

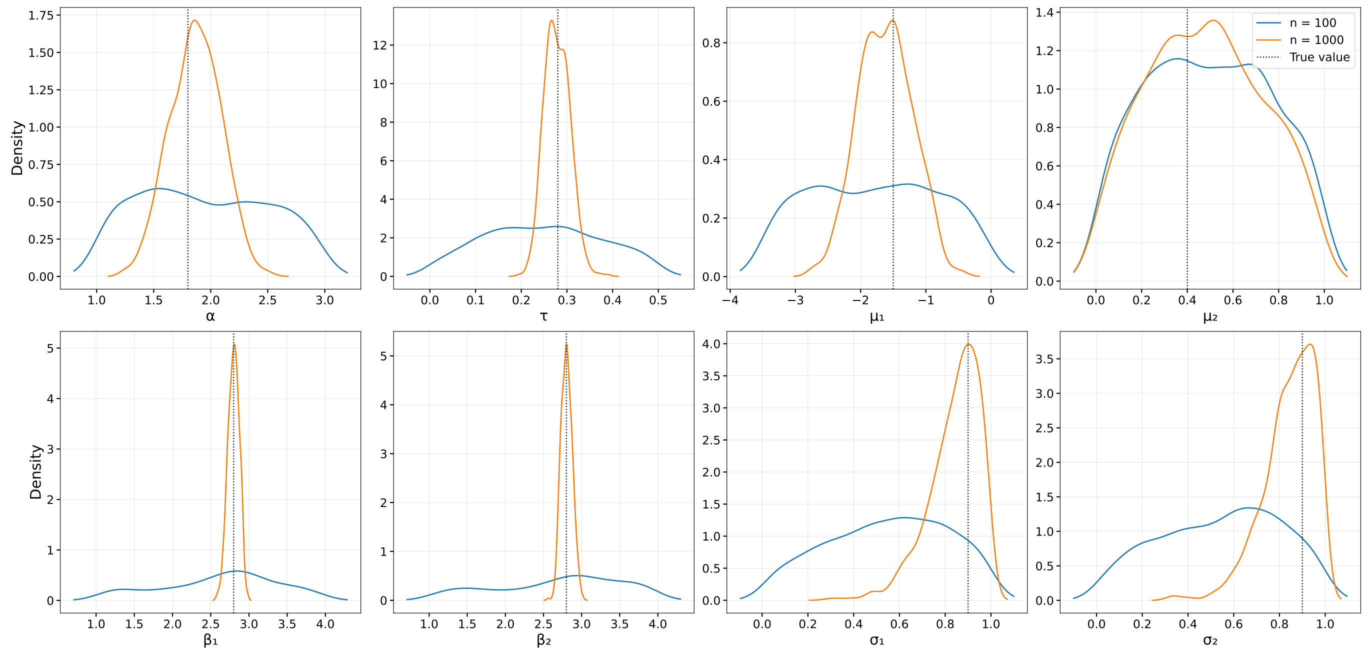

Fig. 9. Structural time series model. Density of parameter estimates using $2p + 1 = 17$ random features.

Simulation-based estimation in statistics. Classical simulation-based approaches include the method of simulated moments (15, 16) and indirect inference (17, 18); see (19) for an overview. (20) extends these ideas by introducing the synthetic likelihood, which is based on the idea of treating user-chosen summary statistics as normally distributed. The method from (20) first simulates many datasets across different parameter values and calculates the summary statistics on each dataset. It then constructs a synthetic likelihood by estimating the mean and covariance of these summary statistics. The parameter estimate is obtained by maximizing this synthetic likelihood. This work inspired an entire literature on synthetic likelihoods (21).

(21) introduce the Bayesian synthetic likelihood approach, which assigns a prior distribution on the parameter and uses Markov chain Monte Carlo methods to sample from the resulting posterior distribution of the parameter conditional on the observed summary statistics. Approximate Bayesian computation is another well-established Bayesian approach, which tries to approximate these posterior distributions by determining whether the simulated summary statistics are below a user-chosen tolerance level (22–24).

Lastly, we point to comprehensive surveys on neural network-based methods. These methods replace the user-chosen summary statistics with representations learned from simulated data–parameter pairs (25, 26). Overall, these simulation-based methods share the idea of substituting unknown quantities with their simulated analogues.

Random features in machine learning. (27) originally proposed using random features to approximate kernel-based predictors. Following up on this work, (28, 29) showed that smooth functions can be well-approximated by a linear combination of randomly chosen basis functions. Since then, researchers have explored using random features for a variety of statistical tasks, such as measuring dependence (30), testing for independence (31), and testing for conditional independence (32).

Embeddings in nonlinear dynamics. (33) initiated the "geometry from a time series" research program, which seeks to reconstruct the attractor of a deterministic dynamical system from an observed signal. They propose concatenating a finite number of lags of the observed signal into a vector, known as a time-delay embedding, and analyzing the resulting space. This idea inspired many subsequent developments in nonlinear dynamics (34).

(35) shows that the time-delay embedding space is topologically equivalent to the underlying p-dimensional state space, once $2p+1$ lags are used for the time-delay embedding. In fact, the minimal embedding dimension is often less, and depends on the box-counting dimension of the attractor (5). (36) shows that a generic set of $2p + 1$ experiments suffices to distinguish two different p-dimensional parameters in analytically parametrized dynamical systems. (37, 38) shows that only $2p+1$ exact measurements at random times are needed for the observability of p-dimensional dynamical systems.

The foundation of these $2p + 1$ results is the classical Whitney embedding theorem (39). We use a probabilistic version of this embedding theorem from (5) to establish the theoretical results for our estimators.

Limit theorems for stochastic processes. Although our theory is based on (extensions of) the physical dependence measure of (1), numerous central limit theorems (CLTs) and laws of large numbers (LLNs) have been established under different regularity conditions. We emphasize that the physical dependence measure conditions we use imply and are implied by β-mixing and strong mixing conditions under certain regularity conditions (2, 3). In what follows, we review existing CLTs and LLNs for mixing processes, mixingales, and Markov chains.

Many CLTs and LLNs have been proven under mixing conditions, which control the decay of temporal dependence so that observations far apart in time become asymptotically independent. For a comprehensive overview on CLTs for strongly mixing processes; see (40) and Chapter 4 of (41). A seminal result by (42) showed that stationary processes satisfying a strong mixing condition admit a CLT. Subsequently, (43) demonstrated stationary ρ-mixing processes satisfy a CLT. Regarding LLNs for mixing processes, (44) and (45) proved strong LLNs for mixing processes (cf. Chapter 3 of (41)).

We now discuss mixingales, which are a generalization of mixing processes and martingale difference sequences. (46, 47) introduced a CLT and LLN for mixingales. Chapter 2 of (48) discusses how martingale difference sequences, strongly mixing sequences, and certain linear processes are all examples of mixingales. We also mention the foundational work of (49), who developed a CLT for stationary processes using a martingale approximation technique.

Next, we discuss Markov chains. (50) provides sufficient conditions for a CLT to hold for Markov chains, and also discusses relationships to results for mixing processes. (51) prove a CLT for reversible Markov chains with a finite limiting variance. (52) provide a CLT for iterated random functions; see (53). Regarding LLNs, (54) shows that a strong LLN holds for Markov chains. For more discussion of CLTs and LLNs for geometrically ergodic Markov chains; see Chapter 17 of (55).

1. WB Wu, Nonlinear system theory: another look at dependence. Proc. Natl. Acad. Sci. 102, 14150–14154 (2005).
2. JB Hill, Mixingale and physical dependence equality with applications. Stat. Probab. Lett. 221 (2025).
3. F Heinrichs, A note on physical dependence and mixing conditions for triangular arrays (2026) arXiv preprint arXiv:2602.04250.
4. J Yorke, W Ott, Prevalence. Bull. Am. Math. Soc. 42, 263–290 (2005).
5. T Sauer, JA Yorke, M Casdagli, Embedology. J. Stat. Phys. 65, 579–616 (1991).
6. BR Hunt, T Sauer, JA Yorke, Prevalence: a translation-invariant almost every on infinite-dimensional spaces. Bull. Am. Math. Soc. 27, 217–238 (1992).
7. H Whitney, Analytic extensions of differentiable functions defined in closed sets. Transactions Am. Math. Soc. 36, 63–89 (1934).
8. R Dahlhaus, Fitting time series models to nonstationary processes. The Annals Stat. 25, 1–37 (1997).
9. R Dahlhaus, Locally stationary processes. Handb. Stat. 30, 351–413 (2012).
10. A Pakes, D Pollard, Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057 (1989).
11. F Mies, A Steland, Sequential gaussian approximation for nonstationary time series in high dimensions. Bernoulli 29, 3114–3140 (2023).
12. F Mies, Strong gaussian approximations with random multipliers. Work. on Stoch. Model. Stat. their Appl. pp. 3–18 (2024).
13. GD Rayner, HL MacGillivray, Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. Stat. Comput. 12, 57–75 (2002).
14. D Prangle, gk: An r package for the g-and-k and generalised g-and-h distributions (2017) arXiv preprint arXiv:1706.06889.
15. D McFadden, A method of simulated moments for estimation of discrete response models without numerical integration. Econom. J. Econom. Soc. 57, 995–1026 (1989).
16. BS Lee, BF Ingram, Simulation estimation of time-series models. J. Econom. 47, 197–205 (1991).
17. C Gourieroux, A Monfort, E Renault, Indirect inference. J. Appl. Econom. 8, S85–S118 (1993).
18. DT Frazier, B Koo, Indirect inference for locally stationary models. J. Econom. 223, 1–27 (2021).
19. C Gourieroux, A Monfort, Simulation-based Econometric Methods. (Oxford University Press, Oxford, England), (1996).
20. SN Wood, Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466, 1102–1104 (2010).
21. LF Price, CC Drovandi, A Lee, DJ Nott, Bayesian synthetic likelihood. J. Comput. Graph. Stat. 27, 1–11 (2018).
22. MA Beaumont, W Zhang, DJ Balding, Approximate bayesian computation in population genetics. Genetics 162, 2025–2035 (2002).
23. MA Beaumont, Approximate bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41, 379–406 (2010).
24. MA Beaumont, Approximate bayesian computation. Annu. Rev. Stat. its Appl. 6, 379–403 (2019).
25. K Cranmer, J Brehmer, G Louppe, The frontier of simulation-based inference. Proc. Natl. Acad. Sci. 117, 30055–30062 (2020).
26. A Zammit-Mangion, M Sainsbury-Dale, R Huser, Neural methods for amortized inference. Annu. Rev. Stat. Its Appl. 12, 311–335 (2025).
27. A Rahimi, B Recht, Random features for large-scale kernel machines. Adv. Neural Inf. Process. Syst. 20 (2007).
28. A Rahimi, B Recht, Uniform approximation of functions with random bases. Annu. Allerton Conf. on Commun. Control. Comput. 46, 555–561 (2008).
29. A Rahimi, B Recht, Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. Adv. Neural Inf. Process. Syst. 21 (2008).
30. D Lopez-Paz, P Hennig, B Schölkopf, The randomized dependence coefficient. Adv. Neural Inf. Process. Syst. 26 (2013).
31. Q Zhang, S Filippi, A Gretton, D Sejdinovic, Large-scale kernel methods for independence testing. Stat. Comput. 28, 113–130 (2018).
32. EV Strobl, K Zhang, S Visweswaran, Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. J. Causal Inference 7 (2019).
33. NH Packard, JP Crutchfield, JD Farmer, RS Shaw, Geometry from a time series. Phys. Rev. Lett. 45, 712–716 (1980).
34. H Kantz, T Schreiber, Nonlinear Time Series Analysis. (Cambridge University Press, Cambridge, England), (2003).
35. F Takens, Detecting strange attractors in turbulence. Symp. on Dyn. Syst. Turbul. Warwick 1980 pp. 366–381 (1981).
36. ED Sontag, For differential equations with r parameters, $2r+1$ experiments are enough for identification. J. Nonlinear Sci. 12, 553–583 (2003).
37. D Aeyels, Generic observability of differentiable systems. SIAM J. on Control. Optim. 19, 595–603 (1981).
38. D Aeyels, On the number of samples necessary to achieve observability. Syst. Control. Lett. 1, 92–94 (1981).
39. H Whitney, Differentiable manifolds. Annals Math. 37, 645–680 (1936).
40. RC Bradley, Basic properties of strong mixing conditions. a survey and some open questions. Probab. Surv. 2, 107–144 (2005).
41. E Rio, Asymptotic Theory of Weakly Dependent Random Processes. (Springer, New York, NY), (2017).
42. M Rosenblatt, A central limit theorem and a strong mixing condition. Proc. Natl. Acad. Sci. 42, 43–47 (1956).
43. IA Ibragimov, A note on the central limit theorem for dependent random variables. Theory Probab. Its Appl. 20, 135–141 (1975).
44. JR Blum, DL Hanson, LH Koopmans, On the strong law of large numbers for a class of stochastic processes. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 2, 1–11 (1963).
45. H Berbee, Convergence rates in the strong law for bounded mixing sequences. Probab. Theory Relat. Fields 74, 255–270 (1987).
46. DL McLeish, Dependent central limit theorems and invariance principles. Annals Probab. 2, 620–628 (1974).
47. DL McLeish, A maximal inequality and dependent strong laws. Annals Probab. 3, 829–839 (1975).
48. P Hall, CC Heyde, Martingale Limit Theory and its Application. (Academic Press, New York, NY), (1980).
49. MI Gordin, The central limit theorem for stationary processes. Sov. Math. Doklady 10, 1174–1176 (1969).
50. GL Jones, On the markov chain central limit theorem. Probab. Surv. 1, 299–320 (2004).
51. C Kipnis, SRS Varadhan, Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. Commun. Math. Phys. 104, 1–19 (1986).
52. WB Wu, X Shao, Limit theorems for iterated random functions. J. Appl. Probab. 41, 425–436 (2004).
53. P Diaconis, D Freedman, Iterated random functions. SIAM Rev. 41, 45–76 (1999).
54. L Breiman, The strong law of large numbers for a class of markov chains. The Annals Math. Stat. 31, 801–803 (1960).
55. SP Meyn, RL Tweedie, Markov Chains and Stochastic Stability. (Cambridge University Press, Cambridge, England), 2 edition, (2009).