Editors' Suggestion

Expected correlation in time-series analysis

Theodore MacMillan,[1] James P. Hilditch,[2] and Nicholas T. Ouellette[1]
[1]*Department of Civil and Environmental Engineering, Stanford University, Stanford, California 94305, USA*
[2]*Department of Earth System Science, Stanford University, Stanford, California 94305, USA*

(Received 26 August 2024; revised 19 November 2024; accepted 7 January 2025; published 14 February 2025)

Time-series analysis often involves the characterization of order or predictability, qualities that are related to internal structure and autocorrelation. Investigating a recently proposed algorithm for solving a density prediction task, we demonstrate that if the same system can be viewed on multiple time scales, there is an inevitable degree of expected order and predictability that increases as the system size grows. In particular, we introduce bounds on the expected second-order structure function and autocorrelation function of a time series where multiple observation scales are available, and conclude with a lower bound on the expected correlation time. Such a lower bound shows that there is an inevitable degree of correlation induced when time-series data is aggregated, quantifying a previously overlooked source of bias towards high correlations.

DOI: 10.1103/PhysRevE.111.024121

I. INTRODUCTION

It is generally expected that the further into the future one is asked to predict, the larger one's error will be. This follows from a general notion of persistence: even if a phenomenon changes in time, it is likely to stay the same over small enough time scales. And if this is not the case and the phenomenon changes rapidly on small time scales, it is generally expected that at least this variability will persist over small time scales. Thus, predictability is usually a decreasing function of time, for example as in weather forecasting, where scales such as the Lyapunov time offer guidance as to how unpredictable a dynamical system will become after a given amount of time [1].

Our expectations change when we are instead asked about the statistical properties of systems at larger times. It would likely be easier to predict the average temperature over the course of a year, for example, than to predict the average temperature for some single month in that year. At the same time, it would be easier still to predict the average temperature over the next five minutes. Predicting averages in dynamical systems is thus not a strictly decreasing function of time.

In the statistical sense, then, we can ask which is harder to predict: short-time averages or long-time averages? Or perhaps intermediate-time averages? To make these questions more concrete, consider the following "density-prediction" game, first introduced by [2]: we are given a sequence of length T comprised of 0 s and 1 s. We are asked to predict the average value of some contiguous subsequence of length w (which will be the "density" of the subsequence) and are allowed to pick the time at which we make our prediction as well as the length of time over which we predict. Drucker [2] and later Qiao and Valiant [3] devised the following prediction scheme: pick t uniformly at random from $\{0, 1, ..., T\}$ and w uniformly at random from $\{2^0, 2^1, ..., 2^{\log T}\}$. Once time t occurs, we predict that the average of the next w entries is equal to the average of the most recent w entries. Surprisingly, this algorithm achieves $\mathcal{O}(\frac{1}{\log T})$ squared error when averaged over their random choice of prediction time t and window w—and thus goes to zero as the sequence length T increases.

This counterintuitive result—that averages over most times can be predicted perfectly only given past information—follows from the following fact about a time series: the average value of a bounded system can vary on some time scales at some times, but not simultaneously on all time scales at all times. In the context of images, a similar observation was noted by Feige [4]:

> Consider... a very large black and white checkerboard pattern. Viewed from a large distance, one pixel in the image will average the value of many checkerboard squares, and hence the image may be uniformly gray (very smooth). Viewed from a very short distance, every square may correspond to many pixels, and then nearby pixels will have the same value, so the image will be very smooth almost everywhere (except on the boundary between squares). However, at some intermediate scale, each square will occupy a small number of pixels (say one pixel, or four pixels), and then adjacent pixels will have very different values and the image will not be considered smooth...

In this paper, we explore this idea in the context of predictability and correlation in time series. We begin by considering the consequences of the density-prediction algorithm of Drucker [2] and prove a similar result for continuous signals. We show that the continuous version of their algorithm ultimately provides several bounds on the expected structure and autocorrelation of a time series, specifically when multiple scales can be considered at the same time. From these bounds we find that the smallest expected correlation time τ_c of a time series grows as $\mathcal{O}((BT)^{\sigma/2})$, where $1/B$ is the smallest time scale available for measurement, T is the largest, and $\sigma < 1$ is a value related to the energy of the time series at a random filter scale. We thus find a quantitative bias toward large autocorrelation values when time series data is averaged, and therefore offers a critical tool in assessing the observed predictability of time series and dynamical systems, allowing researchers to decompose measured correlation into what is expected and what is anomalous.

II. CONTINUOUS DENSITY PREDICTION

Consider a modified density-prediction game: we are given some continuous signal $f(t)$ and are asked to predict its time average, allowed as before to choose the window over which we average and the time at which we predict. Suppose the signal is either bounded such that $|f(t)| < k$ or that we know its average energy $\overline{E} = \frac{1}{T}\int_0^T f(t)^2 dt$, and that there is some minimum time scale $1/B$ over which we can make a prediction. Further, we assume that the signal is periodic so as to ignore boundary effects and to ensure that a Fourier transform exists, noting that the introduction of boundaries will at most introduce a constant error into the resulting bounds [3].

Mirroring the algorithm of Drucker [2], we pick the prediction time t uniformly at random from between 0 and T, and the prediction window $w = e^x$ uniformly at random over values of x between $-\ln B$ and $\ln \frac{T}{2}$. We then predict that the future average value of f between t and $t + w$ is equal to its past average value between $t - w$ and t.

Now, fix a window $w = e^x$. Then our prediction at each time t is the convolution of a rectangle function between $-w$ and 0 (call this $g_{w,\text{past}}$) and the function f. So $f_{w,\text{pred}} = f * g_{w,\text{past}}$, where $*$ denotes convolution. Similarly, the actual value is f convolved with a rectangle between 0 and w, or $f_{w,\text{act}} = f * g_{w,\text{fut}}$. So the error of our prediction at each t is just $\theta_w(t) = f * g_{w,\text{fut}} - f * g_{w,\text{past}} = f * h_w$, where $h_w(t)$ is a convolution kernel with a positive rectangle between 0 and e^x and a negative rectangle between $-e^x$ and 0 (normalized to have an area of 1).

The average (in time) squared error of our algorithm is just the energy of this signal $\theta_w(t)$. So, we can write

$$\begin{aligned}\overline{\theta_w^2} &= \frac{1}{T}\int_0^T (f * h_w)^2 dt \\ &= \frac{1}{T}\int_{-\infty}^{\infty} (\mathcal{F}(f * h_w)(s))^2 ds \\ &= \frac{1}{T}\int_{-\infty}^{\infty} F(s)^2 H_w(s)^2 ds,\end{aligned}$$

where $F(s)$ and $H_w(s)$ denote the Fourier transforms of f and h_w, respectively. An expectation over values of x then gives

$$\begin{aligned}\theta^2 &= \frac{1}{T\left(\ln\left(\frac{T}{2}\right) + \ln(B)\right)} \int_{-\ln B}^{\ln T/2} \int_{-\infty}^{\infty} F(s)^2 H_w(s)^2 ds dx \\ &= \frac{1}{T\ln\left(\frac{BT}{2}\right)} \int_{-\infty}^{\infty} (F(s))^2 \left(\int_{-\ln B}^{\ln T/2} H_w(s)^2 dx\right) ds. \qquad (1)\end{aligned}$$

If we could remove the frequency dependence of the integral involving x, we would have a value of the expected squared error that just involves the energy of f. Fortunately, we can compute the Fourier transform of h_w and, therefore, this integral

$$\begin{aligned}I &= \int_{-\ln B}^{\ln T/2} H_w(s)^2 dx \\ &= \int_{-\ln B}^{\ln T/2} 4\left(\frac{\sin^2(\pi s e^x)}{\pi s e^x}\right)^2 dx \\ &\leqslant 4\int_{-\infty}^{\infty} \left(\frac{\sin^2(\pi s e^x)}{\pi s e^x}\right)^2 dx \\ &= 4\ln 2. \qquad (2)\end{aligned}$$

Remarkably, then, taking an expectation over window size removes the frequency dependence of the filter (see the Appendix A for more details). This integral does not have a well defined value at $s = 0$, but it is clear that a larger or smaller average (DC) value would not affect the error in the algorithm, as any nonzero average would just shift the prediction and actual signal by the same offset. In further analysis we therefore assume that the time series has zero mean (the only change would be to define the energy as the variance of the signal). We can then evaluate the energy of f by itself and simplify to obtain

$$\begin{aligned}\theta^2 &\leqslant \frac{4\ln 2}{T\ln\left(\frac{BT}{2}\right)} \int_{-\infty}^{\infty} F(s)^2 ds \\ &\leqslant \frac{4\ln 2}{T\ln\left(\frac{BT}{2}\right)} \int_0^T f(t)^2 dt.\end{aligned}$$

Given an upper bound on the average energy $\overline{E}$, we then have

$$\theta^2 \leqslant \frac{4\ln 2}{\ln\left(\frac{BT}{2}\right)}\overline{E}. \qquad (3)$$

If only the maximum value k is known, then its energy $E = \int_0^T f(t)^2 dt \leqslant Tk^2$ and we obtain the less restrictive bound

$$\theta^2 \leqslant \frac{4\ln 2}{\ln\left(\frac{BT}{2}\right)}k^2. \qquad (4)$$

Assuming that the energy grows linearly with the size of the system in Eq. (3), we recover the $\mathcal{O}(\frac{1}{\ln T})$ error of Qiao and Valiant [3], with additional factors related to the minimum prediction time and the energy of the signal.

Next, we simulate the density prediction algorithm on three datasets. We take in total 427 time series used in the literature to benchmark prediction tasks, comprised of real time series taken from weather observations and transformer temperature readings [5,6] as well as synthetic dynamical-systems time series [1]. These data represent signals across various sampling rates and energies that allow us to investigate the tightness of this error bound on the prediction algorithm. As discussed before, while not strictly necessary, we transform the time series to have zero mean. To compute the error explicitly, we select a finite sample of window sizes (logarithmically spaced) between 1 and $T/2$, and convolve the time series with the prediction error kernel $h_w(t)$. This enables us to avoid a random sampling approach. Figure 1 demonstrates that our bound is indeed tight for all datasets considered. While we do not consider it, this tightness (for all but a few series) suggests that a lower bound on the error might also be found in addition to our upper bound.

Noncontiguous subsequences

Equation (3) implies that as the number of scales in the system grows, the error of predicting a moving average (over a randomly chosen exponential window) goes to zero. It is

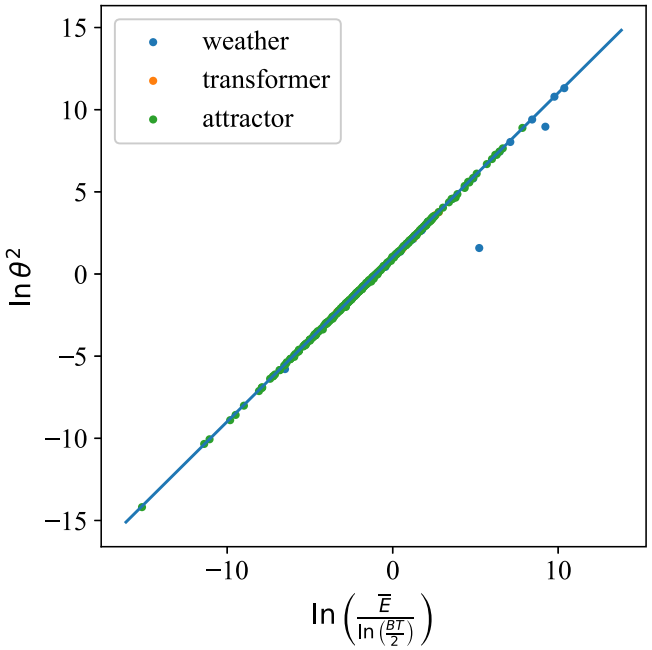

FIG. 1. Average squared error against the energy of the signal scaled by $\ln BT$. The upper bound provided by Eq. (3) is shown for reference.

worthwhile to explore a generalization of this bound, and in particular whether or not the prediction window (rectangle going into the future) and calibration window (rectangle going into the past) must be contiguous.

Our original convolution kernel h_w can be rewritten as $\frac{1}{w}\Pi_w * (\delta(w/2) - \delta(-w/2))$, a rectangle of width w convolved with delta functions of opposite sign centered at $w/2$ and $-w/2$, respectively. If instead of asking for the average temperature of next month, for example, we want to predict the average temperature of a month at some time in the future, the convolution kernel can be modified to $h_{w,\tau} = \frac{1}{w}\Pi_w * (\delta(\tau/2) - \delta(-\tau/2))$. We can rewrite this as $h_{w,\tau} = \frac{1}{w}\Pi_w * (\delta(w\tau/2) - \delta(-w\tau/2))$, implying that we are making our predictions at some multiple of the prediction window, τ now representing the number of multiples into the future. As before, taking $w = e^x$, we get the same initial expression as in Eq. (1), but the integral now is bounded as

$$\begin{aligned} I &= \int_{-\ln B}^{\ln \frac{T}{\tau+1}} H_{w,\tau}(s)^2 dx \\ &= \int_{-\ln B}^{\ln \frac{T}{\tau+1}} 4\left(\frac{\sin(\pi s e^x)\sin(\pi\tau s e^x)}{\pi s e^x}\right)^2 dx \\ &\leqslant 4\int_{-\infty}^{\infty}\left(\frac{\sin(\pi s e^x)\sin(\pi\tau s e^x)}{\pi s e^x}\right)^2 dx. \end{aligned} \tag{5}$$

The value of this integral will provide us with several useful bounds, and can be computed exactly as

$$\begin{aligned} J(\tau) &= 4\int_{-\infty}^{\infty}\left(\frac{\sin(\pi s e^x)\sin(\pi\tau s e^x)}{\pi s e^x}\right)^2 dx \\ &= (\tau-1)^2\ln(\tau-1) - 2\tau^2\ln(\tau) \\ &\quad + (\tau+1)^2\ln(\tau+1), \end{aligned} \tag{6}$$

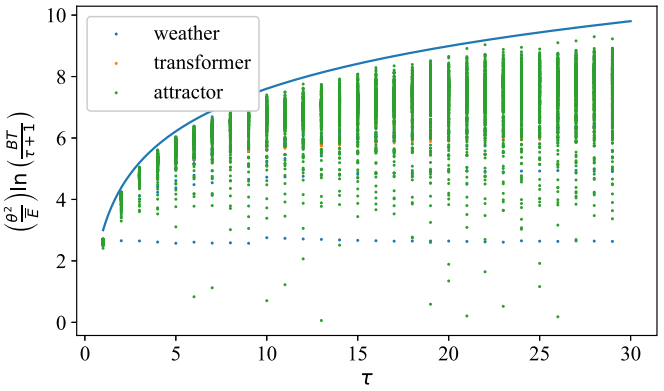

FIG. 2. Average squared error scaled by system size and energy with the upper bound given by Eq. (8).

where a detailed derivation has been provided in the Appendix A. Thus, the more general bound becomes

$$\theta^2(\tau) \leqslant \frac{J(\tau)}{\ln\left(\frac{BT}{\tau+1}\right)}\overline{E}. \tag{7}$$

The number of scales available for prediction is restricted to $\frac{BT}{\tau+1}$ to ensure the prediction and calibration windows do not overlap. When these two windows are contiguous, $\tau = 1$ and $J(1) = 4\ln 2$, reducing this more general bound to that in Eq. (3). As we also show in the Appendix A, we have an upper bound of $J(\tau) \leqslant 3 + 2\ln\tau$, which means we can write the simpler

$$\theta^2(\tau) \leqslant \frac{3 + 2\ln\tau}{\ln\left(\frac{BT}{\tau+1}\right)}\overline{E}. \tag{8}$$

For each value of τ between 1 and 30 (recall that τ is a multiple of the prediction window), we compute the expected error following our modified algorithm for the same datasets previously considered. The results, shown in Fig. 2, demonstrate that our bound is no longer tight for all datasets but that the upper bound is well captured.

III. EXPECTED CORRELATION

The continuous density prediction game has additional relevance outside of guessing temperatures. Consider some phenomenon with a smallest time scale $1/B$ and largest time scale T. A measurement of the system will naturally fall somewhere between $1/B$ and T, with the total number of available scales $\ln BT$ [7]. Often, the first statistical test one would compute for such data would be its autocorrelation, a function critical in producing models of the data [8] and given by

$$G(p) = \frac{1}{T}\int_0^T f(t+p)f(t)dt. \tag{9}$$

The related structure function [9] is given as

$$S(p) = \frac{1}{T}\int_0^T (f(t+p) - f(t))^2 dt, \tag{10}$$

which is often interpreted as measuring the energy of scales less than p (its Fourier transform is the power spectrum).

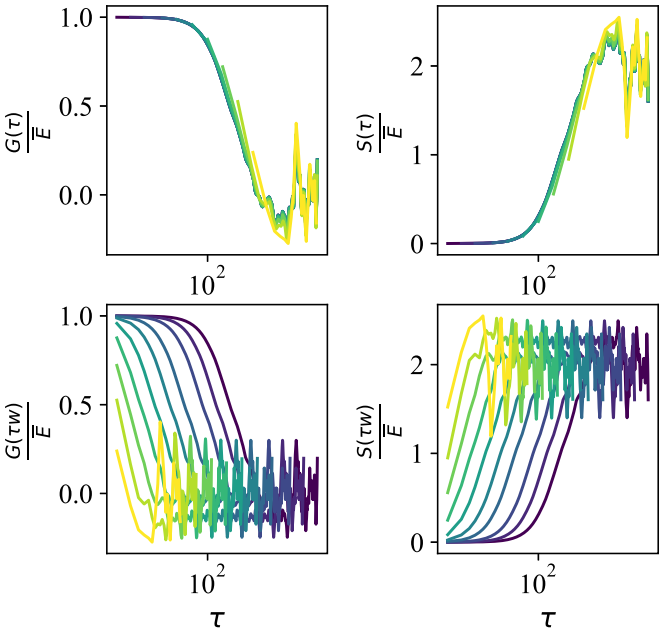

FIG. 3. (Top left) Correlation function and (top right) structure function at different filter scales on the same time axis. (Bottom left) Correlation function and (bottom right) structure function on a per-filter time axis. Increasing brightness corresponds to large scale averaging.

Assuming stationarity, the two functions are related by [10]

$$S(p) = 2G(0) - 2G(p). \tag{11}$$

As we will now demonstrate, the algorithm for predicting running averages suggests a lower bound for the autocorrelation of an averaged time series, provided that several scales are available over which it may be observed. This situation corresponds roughly to the following scenario: there is some natural phenomenon (like those described above) and one is *a priori* unclear as to its intrinsic time scales or is methodologically bound by instrument resolution or storage or simulation size. Thus, the choice of averaging may be treated as random and independent of the smallest scale $1/B$.

If we first measure the signal at scale w, the structure function we would compute (where τ represents multiples of w) is

$$S_w(\tau) = \frac{1}{T}\int_0^T (\overline{f}^w(t+\tau w) - \overline{f}^w(t))^2 dt, \tag{12}$$

and the corresponding filtered autocorrelation is

$$G_w(\tau) = \frac{1}{T}\int_0^T (\overline{f}^w(t+\tau w)\overline{f}^w(t))^2 dt, \tag{13}$$

where $\overline{f}^w$ is the signal f filtered at scale w. Is is useful here to briefly remark on the subtlety of scaling these functions by w instead of shifting strictly by τ. Figure 3 shows the filtered structure functions and autocorrelations scaled and unscaled by the filter size from an example time series (the pressure signal from the weather dataset [5]).

The top row in Fig. 3 shows the functions when they are "properly" scaled, but our further analysis will focus instead on the bottom row of "improperly" scaled autocorrelation and structure functions. The reason for this choice is that if a single resolution is chosen for measurement or simulation, autocorrelation and structure functions are likely to be reported in units of this measurement (or larger). Taking an expectation over filter scales in that case, we look at the expected autocorrelation or structure function scaled by the window over which we are averaging.

This formulation of the expected structure function is then exactly the object treated in the previous section. To see this, observe that

$$\begin{aligned} S_w(\tau) &= \frac{1}{T}\int_0^T (\overline{f}^w(t+\tau w) - \overline{f}^w(t))^2 dt \\ &= \frac{1}{T}\int_0^T \left(\frac{1}{w}\Pi_w * (\delta_{w\tau} - \delta_0) * f\right)^2 dt, \end{aligned}$$

so that its expected value (averaged over possible observation scales) is exactly the expected mean-squared error of the prediction algorithm given in the previous section. The averaged structure function is therefore upper bounded by

$$\langle S(\tau)\rangle \leqslant \frac{3 + 2\ln\tau}{\ln\left(\frac{BT}{\tau+1}\right)}\overline{E}, \tag{14}$$

where the brackets correspond to scale averaging (i.e. $\langle S(\tau)\rangle = \mathbb{E}_w[S_w(\tau)]$). From the identity Eq. (11), we can derive a lower bound on the autocorrelation function (see Appendix B), given by

$$\langle G(\tau)\rangle \geqslant \langle\overline{E}\rangle - \frac{\frac{3}{2} + \ln\tau}{\ln\left(\frac{BT}{\tau+1}\right)}\overline{E}, \tag{15}$$

where the energy is now averaged over observation windows as well as over time. From an autocorrelation function, one can compute several metrics about the underlying time series. One such metric is the correlation time, and one way to define it is the first zero crossing of the autocorrelation function. Given our lower bound, we can also compute a lower bound on the correlation time τ_c as

$$\tau_c \geqslant (BT)^{\frac{\sigma}{1+\sigma}} e^{\frac{-3}{2(\sigma+1)}} - 1, \tag{16}$$

where $\sigma = \langle\overline{E}\rangle/\overline{E} \leqslant 1$ is the ratio of the average filtered energy to the actual energy. Although this ratio depends on τ implicitly as the number of scales we can average over is $\frac{BT}{\tau+1}$, we can use the value at $\tau = 1$ since a larger τ will only lead to a larger σ; $\langle\overline{E}\rangle$ is therefore taken as an average over scales between B and $T/2$. For practical purposes and to isolate the effects of σ, we can obtain the more concise

$$\tau_c \geqslant e^{-3/4}(BT)^{\sigma/2} - 1. \tag{17}$$

For each time series in our datasets, we filtered at increasing scales and computed the first zero crossing of the associated autocorrelation function and averaged across scales. The results are plotted along with lower bounds in Fig. 4.

IV. CONCLUSION

In this paper we have demonstrated that the upper bound on the expected error of a randomized density prediction

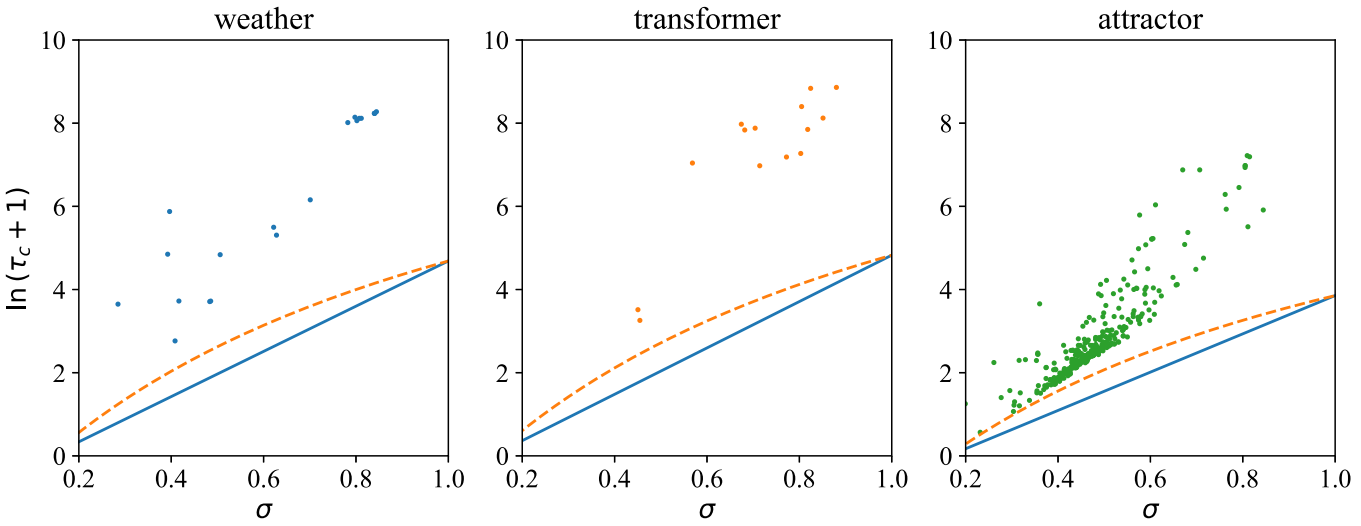

FIG. 4. Expected correlation time against σ. The solid line is the lower bound given by Eq. (17), while the dashed line is the lower bound given by Eq. (16).

algorithm [2,3], when appropriately extended to continuous signals, is mathematically equivalent to a lower bound on the expected correlation function of a randomly averaged time series, ultimately providing a lower bound on the expected correlation time. In doing so, we describe a source of bias toward high correlation when time-series data is aggregated, and offer concrete values against which to compare real-life correlation functions [Eq. (15)], structure functions [Eq. (14)], and correlation times [Eq. (17)].

For reasons unrelated to actual measurement, time series may undergo averaging: for example, due to device storage reasons, data is often aggregated before being saved. Our results show that this averaging procedure must be treated carefully with respect to the final interpretation of results, as the more choices available for averaging, the higher the expected correlation. Interesting extensions of our work would be to look at the expected values of a larger range of statistical tests when multiple scales can be considered. Another interesting avenue would be to search for real-world confirmation of the result in Eq. (17), to see if most reported correlation times fall above or below their expected lower bound. Most correlation times falling below our bound would imply that the choice of averaging is not really independent of the series being observed—and indeed, we might expect this to be the case, as frequently time series will be reported on an averaging scale that looks "interesting." Nevertheless, challenges remain in such an undertaking, for example in estimating the actual range of scales over which a random scale is taken.

Finally, we comment on a qualitative similarity of our results to the family of proofs known collectively as Ramsey theory, a connection remarked upon by Qiao and Valiant [11] in the algorithmic context. Ramsey theory is the study of the emergence of ordered substructure that appears inevitably as the size of a system grows [12,13]. As it is usually used to treat discrete, combinatorial objects, it has not been explored as thoroughly in the context of continuous dynamical systems. However, persistent emphasis on coherent structures in such systems (e.g., [14]) indicates the fruitfulness of establishing better baselines of order and correlation in dynamical systems at large.

ACKNOWLEDGMENTS

T.M. acknowledges financial support from a Stanford Graduate Fellowship and from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656518.

APPENDIX A: DERIVATION OF $J(\tau)$

1. Value of $J(\tau)$

Let

$$J(\tau) = 4\int_{-\infty}^{\infty}\left(\frac{\sin(\pi s e^x)\sin(\pi\tau s e^x)}{\pi s e^x}\right)^2 dx. \tag{A1}$$

We make the substitution $t = \pi s e^x$ such that

$$J = 4\int_0^{\infty}\frac{\sin^2(t)\sin^2(\tau t)}{t^3}dt. \tag{A2}$$

Integrating by parts twice, we have

$$J = 2\int_0^{\infty}\frac{1}{t}\frac{\mathrm{d}^2}{dt^2}(\sin^2(t)\sin^2(\tau t))\mathrm{d}t, \tag{A3}$$

where there is no contribution from the boundary terms. Utilizing trigonometric product identities and then computing the derivatives, we have

$$\begin{aligned}&\frac{d^2}{dt^2}(\sin^2(t)\sin^2(\tau t))\\ &\quad= \cos(2t) + \tau^2\cos(2\tau t) - \frac{1}{2}(\tau+1)^2\cos(2(\tau+1)t)\\ &\qquad- \frac{1}{2}(\tau-1)^2\cos(2(\tau-1)t).\end{aligned} \tag{A4}$$

Being careful with the lower limit of integration, J can be decomposed as

$$J = \lim_{a\to 0^+}\Bigg[2\int_a^\infty \frac{\cos(2t)}{t}dt + 2\tau^2\int_a^\infty \frac{\cos(2\tau t)}{t}dt - (\tau+1)^2\int_a^\infty \frac{\cos(2(\tau+1)t)}{t}\mathrm{d}t - (\tau-1)^2\int_a^\infty \frac{\cos(2(\tau+1)t)}{t}dt\Bigg]. \tag{A5}$$

All these integrals are of the form

$$\begin{aligned}\int_a^\infty \frac{\cos(2\lambda t)}{t}dt &= \int_{2\lambda a}^\infty \frac{\cos(t')}{t'}dt' \\ &= -\mathrm{Ci}(2\lambda a) \\ &= -\gamma - \ln(2a) - \ln(\lambda) + O(a^2),\end{aligned}$$

where we have used the asymptotic expansion of the Cosine integral, Ci, in terms of the Euler-Mascheroni constant γ. When substituting into J there is substantial cancellation, leaving

$$J = \lim_{a\to 0}[-2\tau^2\ln(\tau) + (\tau+1)^2\ln(\tau+1) + (\tau-1)^2\ln(\tau-1) + O(a^2)], \tag{A6}$$

and thus

$$J = (\tau-1)^2\ln(\tau-1) - 2\tau^2\ln(\tau) + (\tau+1)^2\ln(\tau+1). \tag{A7}$$

As required for Eq. (1), the value of $J(1)$ can be computed as

$$\begin{aligned}J(1) &= \lim_{\tau\to 1^+}[\underbrace{(\tau-1)^2\ln(\tau-1)}_{\to 0} - \underbrace{2\tau^2\ln(\tau)}_{\to 0} + \underbrace{(\tau+1)^2\ln(\tau+1)}_{\to 4\ln 2}] \\ &= 4\ln 2.\end{aligned} \tag{A8}$$

2. Upper bound on $J(\tau)$

First we rewrite $J(\tau)$ as

$$J = \tau^2\ln\left(\frac{(\tau-1)(\tau+1)}{\tau^2}\right) + 2\tau\ln\left(\frac{\tau+1}{\tau-1}\right) + (\ln(\tau+1) + \ln(\tau-1)), \tag{A9}$$

and then make use of the limit

$$\lim_{t\to\infty}\left(1+\frac{x}{t}\right)^t = \mathrm{e}^x \iff \lim_{t\to\infty} t\ln\left(1+\frac{x}{t}\right) = x. \tag{A10}$$

The limiting behavior of our terms can then be shown to be

$$\begin{aligned}J &= \underbrace{\tau^2\ln\left(1-\frac{1}{\tau^2}\right)}_{\to -1} + 2\underbrace{(\tau-1)\ln\left(1+\frac{2}{\tau-1}\right)}_{\to 2} + 2\underbrace{\ln\left(\frac{\tau+1}{\tau-1}\right)}_{\to 0} \\ &\quad + 2\ln(\tau) + \underbrace{\ln\left(1+\frac{1}{\tau}\right)}_{\to 0} + \underbrace{\ln\left(1-\frac{1}{\tau}\right)}_{\to 0} \\ &= 3 + 2\ln\tau.\end{aligned} \tag{A11}$$

Thus, as $\tau\to\infty$, $J\to 3+2\ln\tau$. With a little bit more work we can show that this is an upper bound, i.e., $J(\tau) < 3 + 2\ln\tau$. First, we observe that

$$\begin{aligned}\frac{d^2J}{d\tau^2} &= 2[\ln(\tau-1) - 2\ln(\tau) + \ln(\tau+1)] \\ &= 2\ln(1-\tau^{-2}) \\ &< -2\tau^{-2} \\ &= \frac{d^2}{d\tau^2}(3+2\ln\tau).\end{aligned} \tag{A12}$$

Thus, for all $\tau > 1$,

$$\frac{d^2J}{d\tau^2} < \frac{d^2}{d\tau^2}(3+2\ln\tau), \tag{A13}$$

and since they have the same asymptotic behavior as $\tau\to\infty$,

$$\frac{dJ}{d\tau} < \frac{d}{d\tau}(3+2\ln\tau) \tag{A14}$$

and

$$J < 3 + 2\ln\tau. \tag{A15}$$

APPENDIX B: LOWER BOUND ON τ_c

To lower bound the autocorrelation function, we first observe that at any particular filtered scale

$$S_w(\tau) = 2G_w(0) + 2G_w(\tau), \tag{B1}$$

where $S_w(\tau)$ and $G_w(\tau)$ have been defined in the main text. Averaging both sides of the equation over possible filter scales w, we obtain

$$\langle S_w(\tau)\rangle = \langle 2G_w(0)\rangle + \langle 2G_w(\tau)\rangle. \tag{B2}$$

The left hand side of this equation is the expression for which we have our prior upper bound. The first term on the right hand side is the energy of the filtered signal $\overline{E}^w = \int_0^T(\overline{f}^w)^2dt$ averaged over possible filter scales, so $\langle\overline{E}^w\rangle$. The final term is the quantity we want to constrain. So rearranging and dropping the averaging signs on the structure/correlation functions for notational convenience, we find

$$\begin{aligned}G(\tau) &= \langle\overline{E}^w\rangle - \frac{1}{2}S(\tau) \\ &\geqslant \langle\overline{E}\rangle - \frac{\frac{3}{2}+\ln\tau}{\ln\left(\frac{BT}{\tau+1}\right)}\overline{E}.\end{aligned} \tag{B3}$$

The first zero crossing of the autocorrelation is often used as a proxy for the correlation time. So setting $G(\tau) = 0$ and introducing $\sigma = \langle\overline{E}\rangle/\overline{E}$, we have

$$\frac{\frac{3}{2}+\ln\tau_c}{\ln\left(\frac{BT}{\tau_c+1}\right)} \geqslant \sigma, \tag{B4}$$

which after some algebra becomes

$$e^{-3}(BT)^{2\sigma} \leqslant \tau_c^2(\tau_c+1)^{2\sigma} \leqslant (\tau_c+1)^{2+2\sigma}. \quad \text{(B5)}$$

So,

$$\tau_c \geqslant e^{\frac{-3}{2(\sigma+1)}}(BT)^{\frac{\sigma}{1+\sigma}} - 1. \quad \text{(B6)}$$

Since $\sigma \leqslant 1$, a more concise but less tight lower bound is

$$\tau_c \geqslant e^{-3/4}(BT)^{\sigma/2}. \quad \text{(B7)}$$

[1] W. Gilpin, Chaos as an interpretable benchmark for forecasting and data-driven modelling, Adv. Neural Inf. Process. Syst. 1 (2021).

[2] A. Drucker, High-confidence predictions under adversarial uncertainty, ACM Trans. Comput. Theory **5**, 1 (2013).

[3] M. Qiao and G. Valiant, A theory of selective prediction, Proc. Mach. Learn. Res. **99**, 2580 (2019).

[4] U. Feige, Why are images smooth? in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, Rehovot Israel* (ACM, 2015), pp. 229–236.

[5] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, 2021).

[6] J. Xu, J. Wang, M. Long *et al.*, Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, in *Advances in Neural Information Processing Systems 34* (NeurIPS Proceedings, 2021).

[7] If, for example, the smallest scale is $1/B = 10^{-3}$ and the largest $T = 10^3$, the number of available scales (in a base ten sense) would be $\log 10^6 = 6$.

[8] G. E. Box and D. A. Pierce, Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, J. Am. Stat. Assoc. **65**, 1509 (1970).

[9] S. B. Pope, *Turbulent Flows* (Cambridge University Press, Cambridge, 2000).

[10] E. O. Schulz-DuBois and I. Rehberg, Structure function in lieu of correlation function, Appl. Phys. **24**, 323 (1981).

[11] M. Qiao and G. Valiant, Exponential weights algorithms for selective learning, Proc. Mach. Learn. Res. **134**, 1 (2021).

[12] P. Erdös, Some remarks on the theory of graphs, Bull. Am. Math. Soc. **53**, 292 (2016).

[13] R. Graham and J. Spencer, Ramsey theory, Sci. Am. **263**, 112 (1990).

[14] G. Haller, Lagrangian coherent structures, Annu. Rev. Fluid Mech. **47**, 137 (2015).