Quantifying long time memory in phase space trajectories of molecular liquids

Vladimir Ryabov*

School of Systems Information Science, Department of Complex System,
116-2 Kamedanakano-cho, Hakodate-shi, 041-8655 Hakodate, Hokkaido, Japan

Dmitry Nerukh

Unilever Centre for Molecular Sciences Informatics,
Department of Chemistry, Cambridge University, Cambridge CB2 1EW, UK

A trajectory of liquid water simulated using classical molecular dynamics has been analysed in the framework of symbolic dynamics. The behaviour of symbolic subsequences (words) of nine symbols long has been studied at very long times of $1\mu s$. Contrary to naive expectations, the molecular trajectory behaves very differently compared to both a random signal and a random surrogate with spectral properties identical to the molecular trajectory. The molecular system characteristics resemble those of a chaotic map, the Standard map. We conclude that the most probable reason for deviations from randomness in the molecular system is its deterministic dynamics, in particular, the stickiness of periodic islands in the bulk of chaotic motion.

I. INTRODUCTION

It is commonly believed that memory effects in liquids can be characterized and hence understood by calculating various molecular correlation functions, most often atomistic velocity or coordinate autocorrelation functions. These indicate the memory of no longer than few picoseconds, after which time all processes in simple molecular liquids look purely random. However, it has recently been shown from several different perspectives that molecular systems exhibit memory that lasts significantly longer [1–3]. This is especially true for liquid water that has been shown to exhibit slow relaxation, the phenomenon related to $1/f$ frequency dependence of the energy fluctuation spectrum [4, 5].

Apparently, such a simple statistical measure as time (or ensemble) averaged product of time separated values, the two-point correlations function, is inadequate for representing complex multi-point memory effects in molecular liquids. On the other hand, classical molecular ensembles, being deterministic dynamical systems, can benefit from applying geometrical and statistical methods recently developed in the fields of nonlinear dynamics and information theory.

Considering liquids as nonlinear dynamical systems has the advantage of studying the trajectories in suitably chosen projections of the phase space rather than mere coordinates and velocities of one atom or a restricted set of atoms. The full dimensional phase space trajectory contains, at least theoretically, all possible information about the system. Thus, analysing various projections of the phase space trajectory using statistical methods that have a property of extracting substantial non-random information can potentially reveal hidden memory in liquids.

However, direct application of nonlinear dynamics methods, such as the analysis of dimensions, entropies, various complexity measures, Lyapunov exponents, etc., to the phase space trajectories encounters certain difficulties since most of these methods can deal only with low dimensional systems. It is, therefore, infeasible to study the full phase space trajectories of even small molecular ensembles using these methods, not to mention realistic models of liquid consisting of thousands of atoms. Thus, new information theoretic and statistical methods that can advance the analysis of high-dimensional molecular trajectories are highly desirable.

In this paper we present an approach that analyses molecular trajectories in the framework of symbolic dynamics. In this framework the molecular trajectory is converted into a sequence of symbols from an alphabet consisting of only a few symbols. The resulting symbolic sequence can be analysed using various statistical methods. An important point here is that the trajectory is analysed not as isolated symbols but as a sequence of symbolic strings (words). This ensues the extraction of detailed information from the initially continuous trajectory despite seemingly very coarse grained representation of it with only a few symbols [2].

We have analysed the dynamics of liquid water using such symbolic representation and found surprisingly slow convergence of calculated statistical indicators. The statistics of water time series behave fundamentally different compared to those of a pseudo-random sequences (for example, the digits of the π number) or a random surrogate signal that has the correlation function (and hence the power spectrum) identical to that of water. Moreover, water dynamics resembles the behaviour found in a simple chaotic system, the Chirikov-Taylor or Standard map. Thus, we hypothesise that the origin of such non-random properties of molecular trajectory can be in its deterministically chaotic character.

*Electronic address: `riabov@fun.ac.jp`

II. METHOD

We generate molecular trajectories using classical molecular dynamics. Bulk water (periodic boundary conditions) consisting of 392 SPC [6] molecules was simulated using the GROMACS molecular dynamics [7] package. The temperature of the system was kept constant at 300K using Berendsen [8] thermostat. A sufficient equilibration was performed before collecting data for analysis. The velocity of the hydrogen atom of one of the water molecules was used as a time series for the analysis. At the locations where the velocity pierces the xy plane the points of a two dimensional map (cross section) were generated and used as the original real valued time series for analysis. The analysed trajectory of $1\mu s$ long resulted in approximately $3 \cdot 10^7$ data points.

The real valued two dimensional cross section of the trajectory was then symbolised using the angle variable in the polar coordinate frame. The 2π angle was divided into three equal segments and symbols '0', '1', or '2' were assigned to the data points depending on the segment where the values of the angle fall. Such a method of partitioning the phase space and converting the velocity vector values (real numbers) into symbols from the three symbol alphabet provided asymptotically uniform symbolic sequences, that is those with equal occurrence rates for each symbol.

We would like to note that the choice of an appropriate partition for symbolising the data is not a trivial task. In the ideal case it is desirable to find a partition that is generating, i.e. the one that provides a unique representation of the original real valued time series by a symbolic string. However, generating partitions are known to be difficult to calculate, even in comparatively simple cases of two dimensional maps when the equations of motion are known explicitly [9, 10]. The procedure of obtaining a generating partition becomes increasingly more complicated for higher dimensional systems or when it is necessary to find it directly from experimental data, that is when the equations of motion are not known [10]. Our choice of partition is motivated by our earlier experiments [2] where we utilized the method described in [10] for the analysis of Statistical complexity [11] of water trajectories. The approximation to the generating partition obtained with the method of [10] coincides with the one we use in the present work. Another support for this choice of partition consists in the observation that it leads to the highest possible value of topological entropy, $h_T = \log(3)$, defined by the three symbol alphabet that we used. It is known that improper choice of partition can lead to lower values of topological entropy, at least in the case of low dimensional dynamical systems [12]. It is also worth noticing that our partition produces water trajectory time series indistinguishable from both the noise and the Standard map data in terms of the topological entropy.

Our further analysis concerns the statistical properties of small subsequences (words) appearing in the symbolic
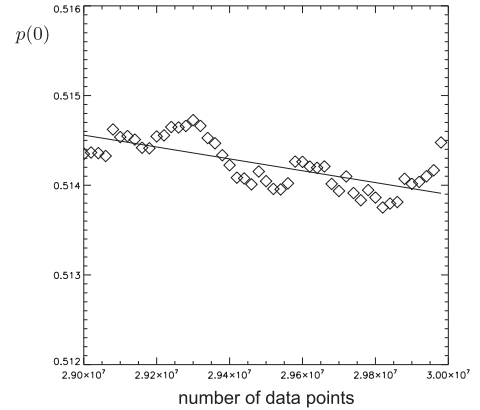

FIG. 1: Convergence of the conditional probability $p(0) = p(0|\{010200100\})$ with the length of time series for an arbitrarily chosen symbolic word '010200100'. The line shows the linear fit to the result of calculating the $p(0)$ value for varying length of the symbolic sequence. The coefficient $\overline{p}(0)$ of the fit defines the slope of the line whereas the mean square deviation of points from the linear fit is denoted as $\sigma_{p(0)}$

sequence obtained from the trajectory. Specifically, we analyse the distributions of the conditional probabilities of the next symbol following the given word

$$p(a_{i+1}|\{a_{i-n}\ldots a_{i-1}a_i\}), \qquad (1)$$

where a_i is the symbol at the time moment t_i, and i varies from n to the number of points in the trajectory. n is the length of the symbolic word $s_i \equiv \{a_{i-n}\ldots a_{i-1}a_i\}$. This parameter has been chosen to be equal to 9 in all the results presented below. The choice of n=9 was defined by the necessity to have high enough number of words for obtaining reliable statistics.

We were interested in the behaviour of (1) for long time trajectories, studying the convergence of the conditional probabilities with time in the limit $t \to \infty$. For this we calculated the slope $\overline{p}$ of the linear trend at the final segment of a long trajectory, as well as the fluctuations σ_p of the average value around the trend line (see Fig. 1).

To compare the results obtained for water simulation with similar characteristics found in other systems with well known properties we have calculated the same statistical indicators for (i) a random signal obtained with a standard random number generator or from a more complicated algorithm generating the digits of the ternary expansion of the number π, (ii) a random surrogate obtained from the molecular signal, and (iii) the two-dimensional sequence of points generated by the Standard map.

As for the random signals, we used pseudo-random sequences generated by specially designed algorithms. Although such signals can be regarded as deterministically chaotic, they are generated by numerical procedures equivalent to high dimensional dynamic systems. Therefore, they can not be easily distinguished in a computer experiment from genuine random sequences obtained by

physical (hardware) noise sources, like, for example, thermal noise generators. The simplest example we consider is randomly shuffled version of the symbolic string obtained from the water data time series. For another example we took the ternary expansion of the number π that, although generated by a deterministic algorithm, can be considered as a high quality noise source due to the complexity of the algorithm that performs the calculation. Both types of random signals led to identical results in terms of distributions of conditional probabilities described here.

The surrogate signal was obtained using a standard technique [13] of transforming the analysed time series into the one with identical power spectrum, but belonging to the class of linear Gaussian processes. The computational procedure in its simplest implementation includes three steps:

- making the Fourier transform of the data;

- randomizing the phases of the Fourier components, keeping the amplitudes intact;

- making the inverse Fourier transform.

The conclusions on the difference between the original data series and a set of computer generated surrogates can be then derived by a suitably chosen discriminating statistics analysis [13].

The Standard map time series were generated with the pair of deterministic equations

$$
\begin{aligned}
P_{n+1} &= P_n + K \sin \theta_n, \\
\theta_{n+1} &= \theta_n + P_n + K \sin \theta_n,
\end{aligned}
$$

where P and θ are computed mod 2π and K is a positive parameter that defines the system's behaviour. An example of a chaotic trajectory in this system at the value of $K = 6.908745$ is shown in Fig. 2, where the dynamics contains a large chaotic area with two stability islands symmetrically located with respect to the origin. Infinite number of smaller islands also exist close to the large ones, but they are not discernible at the picture resolution.

Assigning symbols to the trajectory points was done by dividing the phase space into three segments, such that $a_i \equiv 0$ for $\theta < -\frac{\pi}{3}$, $a_i \equiv 1$ for $-\frac{\pi}{3} \leq \theta < \frac{\pi}{3}$, and $a_i \equiv 2$ for $\theta \geq \frac{\pi}{3}$.

III.   RESULTS AND DISCUSSION

Fig. 3 shows the scatter plot of conditional probabilities (1), their trend and variance for the random signal. The values for symbol '0' are shown in the middle and bottom plots, similar plots for the other two symbols look identical. The high symmetry of all three characteristics can be clearly seen.

The results for the random surrogate obtained from the molecular signal is shown in Fig. 4. Here the statistical correlations produce a distinctive structure on the
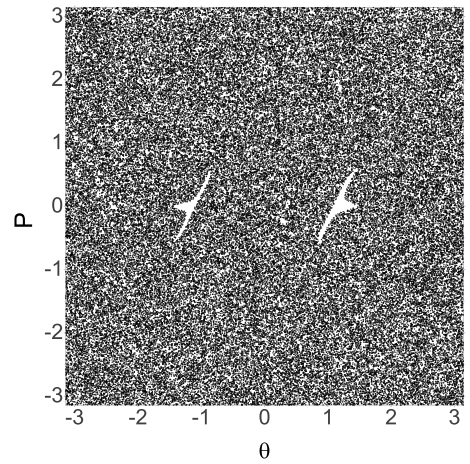

FIG. 2: The Standard map trajectory in chaotic regime

$p(0) - p(1)$ plot. The rotation symmetry of the scatter plot corresponds to the three symbol alphabet used in the analysis. The difference from the simple random behaviour consisting in the appearance of three (approximately Gaussian) clusters of points in the top panel of Fig. 4 instead of a single cluster observed in Fig. 3 is defined by the presence of short and intermediate time correlations in the surrogate or, in the language of the power spectrum, finite equivalent bandwidth of the noise representing the surrogate time series. Large variations in the trend and the variance compared to the uncorrelated random signal are obvious. There is also a tendency of more probable symbolic words to have smaller variance (bottom panel of Fig. 4). Such behaviour can be interpreted as a natural consequence of the statistical property of the estimates to have smaller variance for larger data sets.

A fundamentally different picture presents the molecular system, Fig. 5. Here a much more complex pattern is formed by the conditional probability points. The distributions of the trend and the variance are substantially different from those found for the random signal and the surrogate time series. Also, in contrast to the surrogate signal, there is a small number of words that are substantially more probable than the rest of the sequences. These more probable sequences also have both smaller trend and variance compared to the rest of the sequences.

The results for the Standard map present an extremely different case, Fig. 6. First of all, the conditional probabilities themselves form a very sophisticated pattern in the scatter plot, which even looses its three fold symmetry despite the equal occurrence probability for the symbols in the analysed symbolic data string. The trend and the variance show the variations of several orders of magnitude. We interpret this result as a manifestation of chaotic dynamics that produces highly dispersed values for conditional probabilities and, moreover, lead to poor convergence of their values. Chaotic motion in the Standard map is known to possess segments of sub-

stantially different statistical properties when the trajectory visits the areas close to stability islands. The dynamics becomes 'sticky' in the sense that the trajectory spends a long time in the vicinity of the periodic islands if the dynamics brings the trajectory sufficiently close to one of them. The 'stickiness' of the periodic islands, thus, results in very long convergence times for the conditional probabilities, as well as highly dispersed values of the probabilities themselves presenting a complicated pattern shown in the top panel of Fig. 6.

The observed structures in the scatter plots does not depend on the way we introduce symbolisation of the data. The pattern shown in Fig. 6 has been obtained by introducing a partition of the phase space shown in Fig. 2 into three vertical stripes of equal size. Shifting
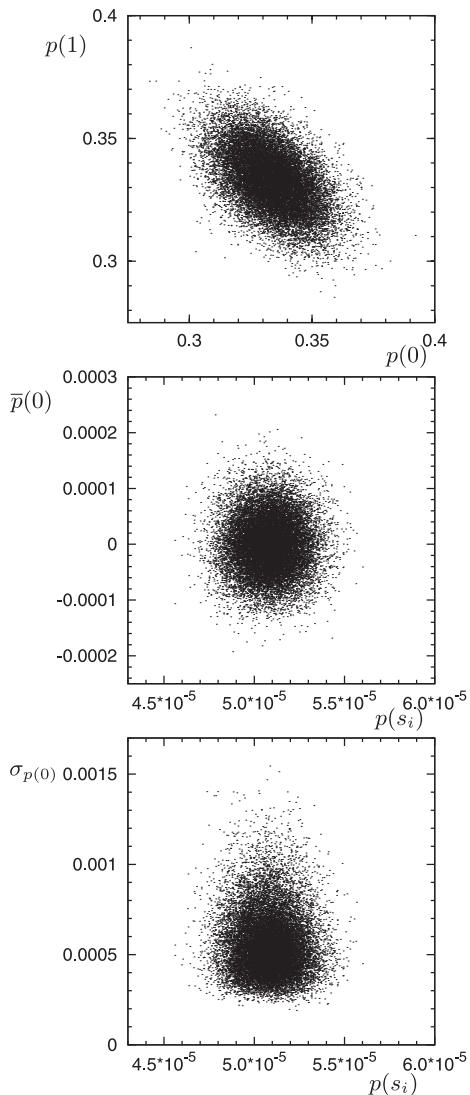
the partition along the θ axis or changing the orientation of partition do not bring qualitative changes into the patterns presented in Fig. 6. For example, in Fig. 7 we plot the same distributions, but for the case of partitioning the phase space into three horizontal stripes of equal size. Note that the shape of the patterns in the scatter plots remains qualitatively same, thus supporting our hypothesis on its origin from the deterministic dynamics of the underlying map.

IV. CONCLUSIONS

We have analysed the behaviour of molecular phase space trajectory in liquid water simulation using classical molecular dynamics, symbolic dynamics, and statistics.

FIG. 3: Conditional probabilities of '0', $p(0|\{s_i\})$, and '1', $p(1|\{s_i\})$ (top), and their trend $\overline{p}(0)$ (middle) and variance $\sigma_{p(0)}$ (bottom) against the word occurrence $p(s_i)$ calculated for the symbolic string of length $3 \cdot 10^7$ for the purely random process (ternary expansion of the number π)
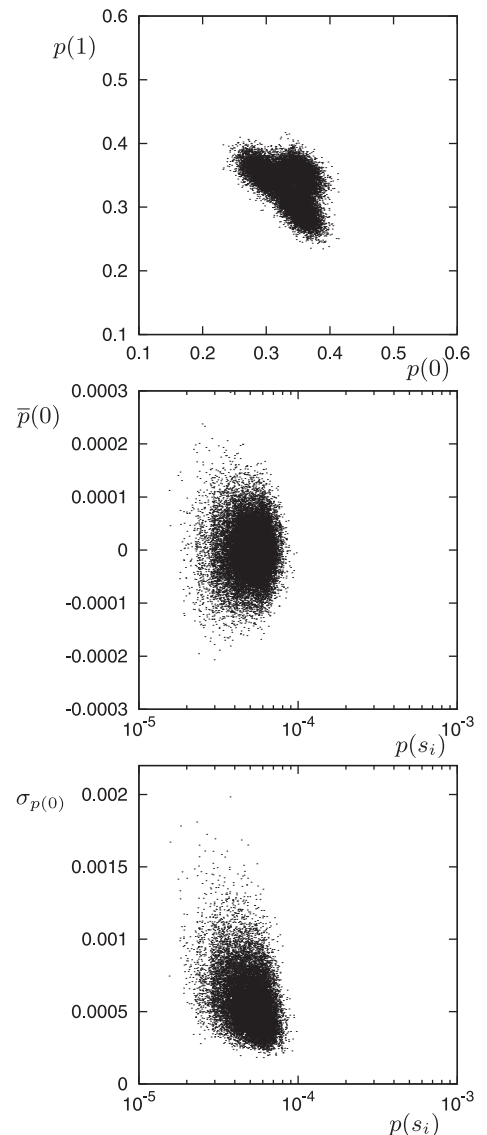

FIG. 4: Same as the previous picture, but calculated for the surrogate time series

Despite the homogeneous nature of the system composed of identical molecules, we have found significant deviations from 'simple' random behaviour at the times of the order of $1\mu s$. On the other hand, we observed certain similarities between the trajectories calculated from the simulation of water dynamics and the dynamics of a classical two dimensional map modelling the kicked rotator (the Standard map). Our statistical approach has been focused on finding the signatures of chaotic dynamics in such a high dimensional dynamical system as the ensemble of interacting molecules. Finally, we came to a conclusion that the statistical characteristics of the trajectories in the molecular system occupy intermediate position between the random surrogate and the chaotic map. We believe that the deviations from randomness in the molecular system are caused by its deterministi-

cally chaotic dynamics that includes 'sticky' areas in the phase space. The stickiness in area preserving (Hamiltonian) maps is a well known property of the dynamics, well-documented for low dimensional systems [14]. There is, however, very few studies of this phenomenon for systems with multiple degrees of freedom.

When the dimensionality of the system becomes large, the transport properties are no longer defined by impenetrable barriers formed by tori, but some essentially new features such as Arnold diffusion emerge as a result of torus break-up. The destroyed tori have complex structure, and some of them are unstable (analogous to saddle points in the case of Standard map). The chaotic trajectories can be trapped by such structures, therefore the observed statistical properties of an arbitrary chaotic trajectory can strongly depend on their presence in the
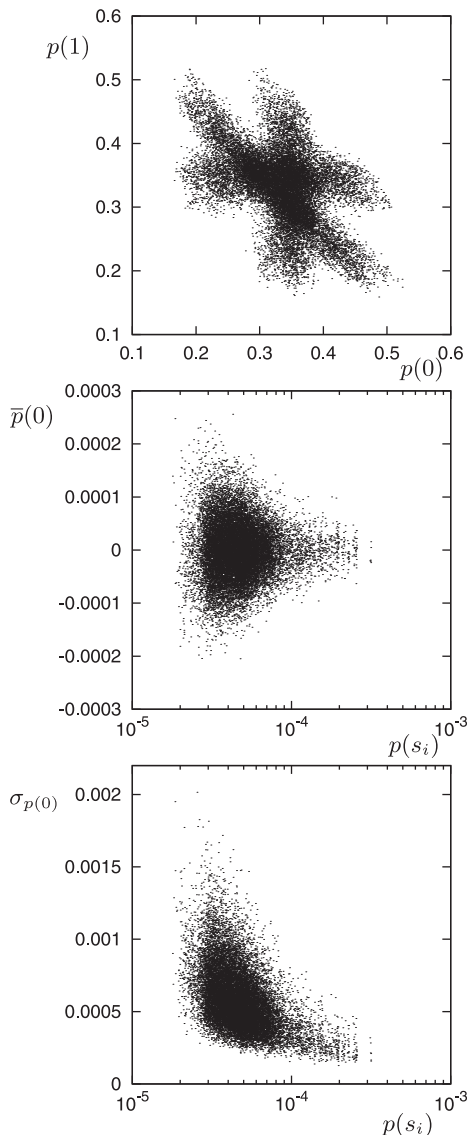

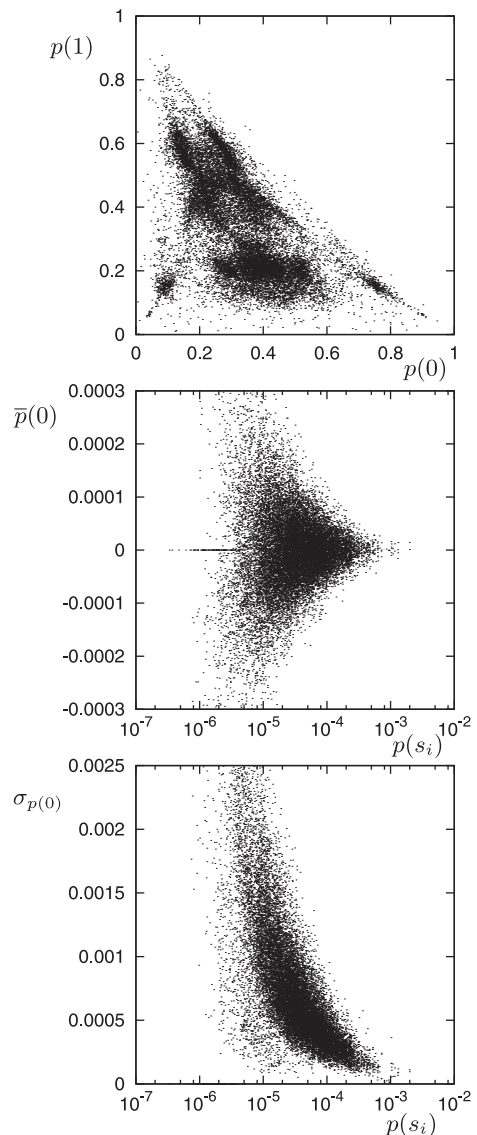FIG. 5: Same as the previous two pictures, but for water signal

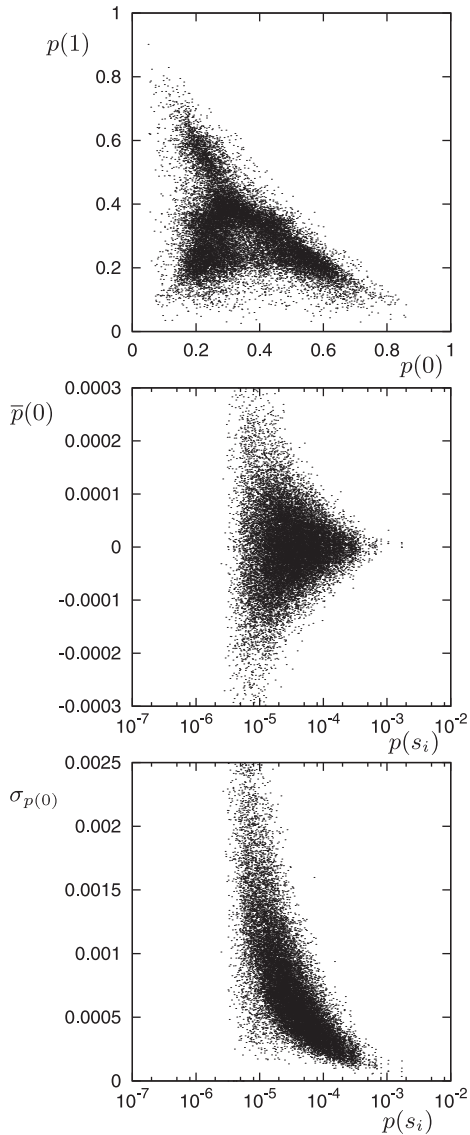FIG. 6: Same as previous three pictures, but for Standard map

FIG. 7: Same as the previous picture, but for different partitioning of the Standard map

phase space. For example, the stickiness has been demonstrated to exist in the Sinai billiard, although the barrier has zero measure in phase space [15]. Moreover, the statistics of recurrence times similar to that found in the Standard map has been recently reported for the case of Arnold diffusion in a 5-dimensional Hamiltonian system [16]. This finding suggests that the properties of Arnold diffusion can be strongly affected by the phenomenon of stickiness.

[1] A. M. Berezhkovskii and G. Sutmann, Phys. Rev. E **65**, 060201 (2002).

[2] D. Nerukh, V. Ryabov, and R. C. Glen, Physical Review E **77**, 036225 (2008).

[3] D. Nerukh, V. Ryabov, and M. Taiji, Physica A: Statistical Mechanics and its Applications **388**, 4719 (2009), ISSN 0378-4371.

[4] M. Sasai, I. Ohmine, and R. Ramaswamy, The Journal of Chemical Physics **96**, 3045 (1992).

[5] A. Shudo, K. Ichiki, and S. Saito, EPL (Europhysics Letters) **73**, 826 (2006).

[6] H. J. C. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (D. Reidel Publishing Company, Dordrecht, 1981), pp. 331–342.

[7] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comp. Chem. **26**, 17011718 (2005).

[8] H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, 1991), p. 139155.

[9] F. Christiansen and A. Politi, Nonlinearity **9**, 1623 (1996).

[10] M. Buhl and M. B. Kennel, Physical Review E **71**, 046213 (2005).

[11] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[12] E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Zyczkowski, Physica D: Nonlinear Phenomena **154**, 259 (2001), ISSN 0167-2789.

[13] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D: Nonlinear Phenomena **58**, 77

(1992), ISSN 0167-2789.

[14] G. M. Zaslavsky, *Hamiltonian Chaos and Fractional Dynamics* (Oxford University Press, 2005).

[15] G. M. Zaslavsky and M. Edelman, Phys. Rev. E **56**, 5310

(1997).

[16] A. Shojiguchi, C.-B. Li, T. Komatsuzaki, and M. Toda, Phys. Rev. E **76**, 056205 (2007).