# POPULATION AND REPLICATE VARIABILITY IN AN EXPONENTIAL GROWTH MODEL*

A. Kleczkowski

Department of Plant Sciences, University of Cambridge
Cambridge CB2 3EA, England

*(Received February 1, 2005)*

*Dedicated to Professor Andrzej Fuliński on the occasion of his 70th birthday*

We have studied variability and predictability of population behaviour in a simple model of exponential growth. Population variability is related to uncertainty of prediction for the dynamics conditioned upon the initial state only. We contrasted it with replicate variability, defined in terms of short-term predictability along a single realisation of a stochastic process. We show that for exponential growth, the population variance increases proportionally to the square of the current population size, whereas the replicate variance is a linear function of the population size. Thus, for large population sizes, the relative predictability for a single population is much better than for an ensemble of realisations. This stands in contrast with the behaviour of a simple stochastic process (Ornstein–Uhlenbeck process), where the population and the replicate variances have similar behaviour. The results have profound consequences for parameter estimation and prediction for many stochastic population models based on the exponential formula.

PACS numbers: 05.45.Ra, 05.45.Xt, 87.23.Cc

## 1. Introduction

In analysis of physical, chemical and biological systems incorporating a stochastic component, we are most often interested in general properties of a (real or hypothetical) ensemble of such systems. Thus, we typically ask a question: What would a typical behaviour be of our biological system, were we to repeat the experiment again? In this approach, each individual realisation of a stochastic process has no separate meaning and we concentrate

* Presented at the XVII Marian Smoluchowski Symposium on Statistical Physics, Zakopane, Poland, September 4–9, 2004.

on the ensemble (or population) properties. However, in practical applications we are often forced to consider a single realisation. We often want to predict how a particular population would evolve in the future, or whether its behaviour is similar to or different from, other historical records. This is particularly important in population biology and epidemiology, where we are faced with analysing and predicting a single outbreak of a disease or a particular invasion of species that takes place in 'real' time.

For systems with ergodic properties, when we can substitute time for replication and replication for time, these two approaches are interchangeable. For highly non-equilibrium processes, like epidemic outbreaks or invasions of species, this is no longer the case. The progression of number of infected individuals as a function of time forms a unique and highly correlated sequence, limited in numbers and duration, as the disease passes through the population and dies out. For such systems there is no steady state and therefore no stationary distribution [1]. Thus, most of the traditionally used analysis tools [2, 3] are not applicable. The outcomes of epidemics may also strongly depend on the initial state and/or substantially vary between individual replicates [4].

Many biological, physical and chemical systems are characterised by large differences between outcomes of replicates within the same experiment as well as by a high sensitivity to small changes in the conditions under which experiments are repeated [5]. This feature makes population (or ensemble) predictions very difficult. It is therefore essential to understand sources of variability, in order to maximise the chances for a successful prediction. A key element of our approach lies in the relationship between variability and prediction. Outcome of processes that are characterised by large variability cannot be predicted with a good confidence. We can therefore use predictability as a measure of variability.

We are considering a simplest model of disease spread through a large (infinite) population, concentrating on two contrasting approaches to variability: *Ensemble* variability is defined as difference between two (or more) observations of an ensemble of populations, sampled at the same time, but regardless of which individual realisation they come from. This forms the 'usual' way in which variability is typically introduced for stochastic systems. We contrast this definition with *replicate* variability, that is associated with a single realisation. As a single realisation of a stochastic process is not a well-defined object, we need to define very carefully what is meant by the replicate variability. We achieve this by analysing predictability along a certain trajectory, conditioned on its history. We have chosen an exponential growth model for its simplicity, but also because it forms the basis for many models of a population behaviour [6].

## 2. Model

The exponential growth model is defined through the following master equation [7]

$$\frac{d}{dt}P\left(x,t\right)=b\left\{\left(x-1\right)P\left(x-1,t\right)-xP\left(x,t\right)\right\}\ , \tag{1}$$

where $P\left(x,t\right)$ is a probability of observing $x$ individuals in a population at a given time $t$, and $b$ is an infection rate. The probability is interpreted here in the 'frequentist' sense, as a proportion of replicate populations in which exactly $x$ individuals are observed at time $t$, assuming an infinite ensemble of replicated populations [9]. The model is analogous to a simple birth model without death, and so the process is not stationary.

Equation (1) can be solved analytically and $P\left(x,t\right)$ is a combination of exponentials of form $\exp\left(-bxt\right)$ [2]. Appendix A gives the details of the solution and its properties [8].

Since no analytical solution can be obtained for the short-term predictability, in the rest of the paper we will concentrate on simulations. For a fixed realisation $k$, the number of new individuals added to the system between time $t$ and $t+dt$ is calculated as a Poisson variable, with a rate given by $bx_k\left(t\right)\,dt$.

$$x_k\left(t+dt\right)=x_k\left(t\right)+\mathcal{P}\left(bx_k\,dt\right)\ . \tag{2}$$

The time step $dt$ is assumed to be small, and we checked the results by varying it over several orders of magnitude, while keeping the overall rate $b\,dt$ constant. The population was initiated from a fixed number of individuals, $x\left(0\right)$, corresponding to $P\left(x,0\right)=\delta\left(x-x\left(0\right)\right)$. The model have been simulated for 2000 steps, assuming $b=0.0025$ and the time step $dt=0.01$ (thus $t=0,\ldots,20$). 2000 replicates were generated for $x\left(0\right)=1$ and $x\left(0\right)=100$.

The *ensemble* variability is quantified by calculating the second central moment (variance) based on the empirical estimate of $P\left(x,t\right)$ obtained using (2). Quantifying the *replicate* variability is much more difficult, as a single replicate $k$ has no meaning in the 'frequentist' interpretation. An autocorrelation function or other similar characteristics cannot be used due to non-stationary character of the series $x\left(t\right)$. Instead, we use the predictability to characterise the uncertainty along each replicate curve.

A replicate $k$ is defined by specifying a finite and discrete set of points $x_k\left(t_n\right)$ along the trajectory, $\left\{t_n;n=1,2,\ldots\right\}$. This can be interpreted in terms of a set of points that has been measured from an outcome of a process. The underlying stochastic process $x_k\left(t\right)$ (which can be either continuous or discrete) may change its value between those measurements, but we assume

that those changes are not directly observable. Hence, a single replicate is properly defined as an equivalence class of all realisations of the stochastic process $x(t)$ agreeing with the given one on a set of measured points $x_k(t_n)$.

We want to predict future values of $x_k(t')$ given $x_k(t_n)$, for $t' > t_n$ ($t_n$ is the $n$-th observed point), for a fixed $k$ and $n$. We are concentrating on the local predictability, so that we assume that $t' - t_n$ is small. We have chosen $t' = t_n + 1 = t_n + 100\,dt$ ($dt = 0.01$). The birth process (1) is Markovian, and so the state of the system $x_k(t_n)$ at $t_n$ contains complete information about the future dynamic. Thus, to predict the values at $t' > t_n$ we use (1) with the initial condition equal to $x_k(t_n)$. After choosing a replicate $k$ at random, we used 2000 simulations to estimate the probability distribution $P(x, t')$ and variance at time $t'$. This procedure was then repeated for all values of $n$.

We also consider an alternative model, for which the increase (or decrease) in the population size is independent of the current value. An Ornstein–Uhlenbeck process is simulated by drawing increments from a normal distribution with zero mean and unit variance, and discounting the previous value of $x$ by a factor $\lambda$. We consider two cases, one with a small correlation between successive steps, corresponding to $\lambda = 0.9$, and with a high correlation, for which $\lambda = 1.0$ (equivalent to the pure diffusion process).

$$x_k(t + dt) = \lambda x_k(t) + \mathcal{N}(0, 1) \ .$$

## 3. Results

The number of infected individuals at a given time, $x_k(t)$, simulated from (1), broadly follows an exponential growth, with a strong increase in the overall ensemble variability over time, Fig. 1(a) (for $x_k(0) = 1$) and Fig. 2(a) (for $x_k(0) = 100$). The mean-field approximation, given by

$$x_{\mathrm{MF}}(t) = x(0) \exp(bt) \ , \tag{3}$$

represents the general trend (thick line in figures 1 and 2), see also Appendix A. However, there is a difference in *relative* levels of variability for different initial conditions. For a small initial value (Fig. 1), there is a large variation around the mean behaviour (note that this effect is relative to the overall mean value). The exponential trend can be removed, by plotting either $\log(x)$ (figure 1(b)) or $\log(x) - bt$ (figure 1(c)) as a function of $t$. The first transformation produces a straight-line relationship with a slope $b$, for large $t$ (for large $x(t)$), whereas the second plot emphasises stochastic component of the dynamics. When a larger initial value is used ($x(0) = 100$), the variability is relatively smaller (figure 2), but the mean value is much larger as well. However, as can be seen by comparing figures 1(c) and 2(c), the *relative* variability is smaller in the case with the larger starting value.

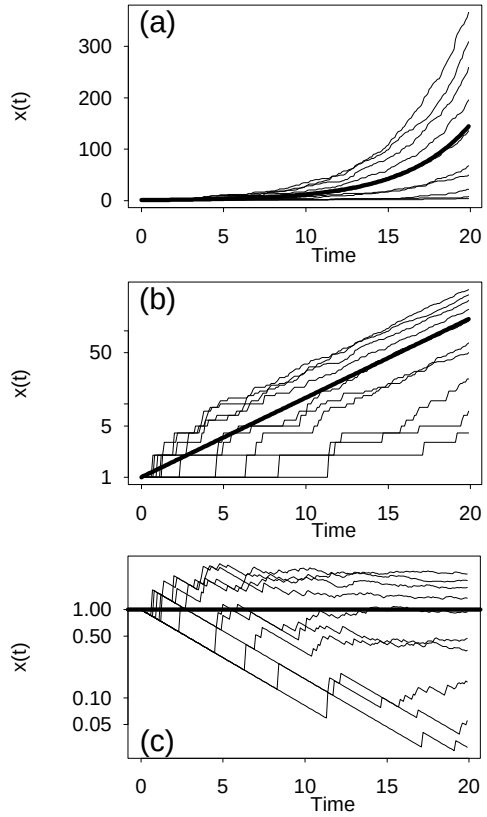


Fig. 1. Examples of trajectories generated for the exponential growth model (1) for small initial condition ($x\,(0) = 1$, 10 replicates are shown, differing only by a different sequence of random numbers in (2)). (a) shows the untransformed values of $x\,(t)$, (b) is the same as (a), but with logarithmic scale on the $y$-axis, whereas in (c), the overall trend given by (3) is removed (trajectories show the ratio of $x\,(t)$ to $x_{\mathrm{MF}}\,(t)$ on a logarithmic scale). The thick line in all plots corresponds to $x_{\mathrm{MF}}\,(t)$.

The difference between the ensemble variability and the replicate variability in figure 1 is a striking feature of the simulation results for $x\,(0) = 1$. If the only information about a given population is its initial size at $t = 0$, the relative uncertainty in the population size at $t \gg 0$ is very large. This stands in contrast to the relative smoothness of individual realisations shown in figure 1(a), particularly for large values of $t$. Removing the exponential trend (figures 1(b) and 1(c)) makes the difference even more apparent. For example, in figure 1(b) (logarithmic transformation), individual trajectories follow a straight-line relationship, each with a similar slope which is asymptotically equal to $b$. Initially (*i.e.* for short times $t$ and small values of $x\,(t)$),

there are large differences both within and among individual realisations, but the within-replicate component declines and the trajectories are becoming relatively smoother as $t \to \infty$. This is particularly clearly visible in figures 1(c) and 2(c).

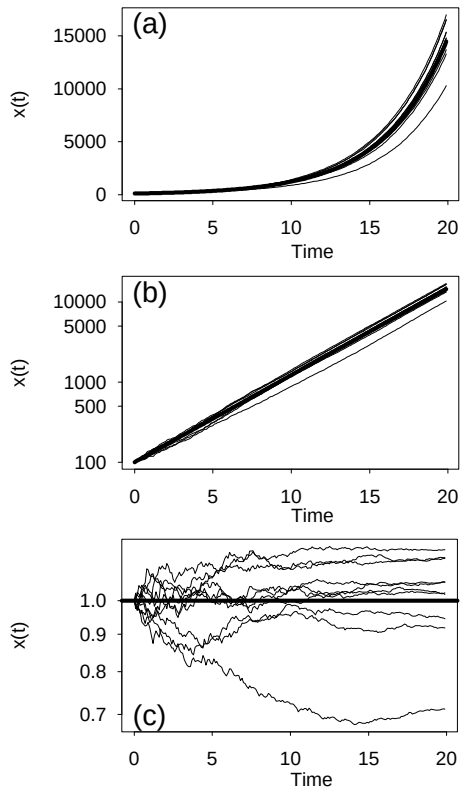


Fig. 2. The same as in figure 1, but for large initial condition ($x\,(x,0) = 100$). (a) shows the untransformed values of $x\,(t)$, (b) is the same as (a), but with logarithmic scale on the $y$-axis, whereas in (c), the overall trend given by (3) is removed (trajectories show the ratio of $x\,(t)$ to $x_{\mathrm{MF}}\,(t)$ on a logarithmic scale). The thick line in all plots corresponds to $x_{\mathrm{MF}}\,(t)$. Note changed range of $y$-axis in comparison to figure 1.

The master equation describes the evolution of the probability distribution corresponding to a behaviour of the whole ensemble of replicates. The ensemble variability can be characterised by computing the second central moment for $P\,(x,t)$, or alternatively by computing a variance (or an interquantile distance), for a given time $t$. We can interpret the variance at any given time $t$ as a measure of uncertainty with which we can predict the size of an epidemic at $t$, conditioned on a fixed initial state at $t = 0$. Dynamic

of variance for $x(0) = 1$ is shown in figure 3. Because the initial distribution is sharp, the variance is initially small, but soon reaches its asymptotic behaviour (*cf.* Appendix A). Afterwards, the variance grows exponentially (figure 3(a); because the $y$-axis is logarithmic, the graph is a straight line), but faster than $x(t)$ (figure 3(b)). Further analysis shows that the population variance behaves asymptotically like $x^2(t)$ (figure 3(b)), so that the standard deviation is asymptotically proportional to $x(t)$ (figure 3(c)), *cf.* Appendix A.

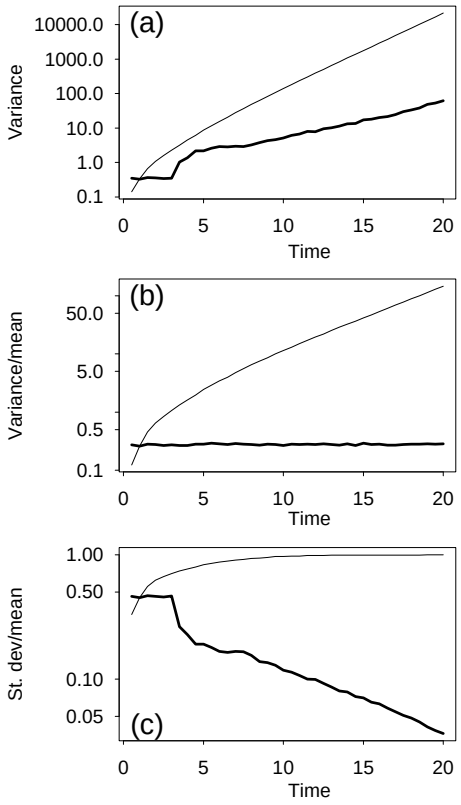


Fig. 3. Population variability (thin lines) is compared with replicate variability (thick lines) for the population model (1) with $x(0) = 1$ (as in figure 1). Variance is shown in (a), variance-to-mean ratio in (b) and standard-deviation-to-mean ratio in (c). Note the logarithmic scale on the $y$-axis. For the replicate analysis a single randomly chosen realisation was analysed.

The single-replicate uncertainty is characterised here by the variance for the probability distribution associated with a short-term prediction conditioned on initial values taken from a randomly chosen replicate at different

times ($t = 1,\ 2,\ \ldots$). In contrast to the ensemble variance, the replicate variance grows much slower with time (figure 3(a)) as it is proportional to $x\,(t)$ rather than $x^2\,(t)$ (figure 3(b)). For small initial values ($x\,(0) = 1$ as in figures 1 and 3), there are substantial differences between estimates of the replicate variance based on different realisations, although the qualitative behaviour is the same (figure 4(a)). The differences are fully linked to different values of $x\,(t)$, reflecting the proportionality of the variance to the average value of $x\,(t)$ for a given $t$, see figure 4(b).

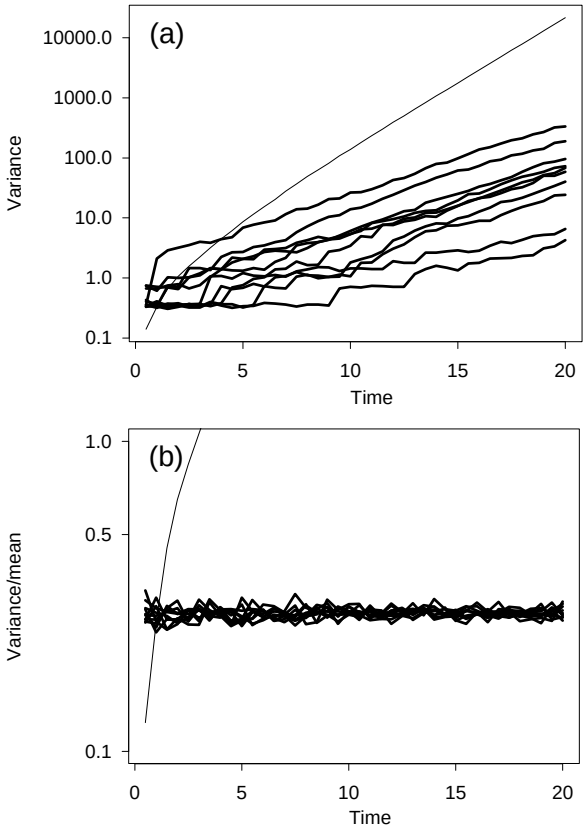


Fig. 4. Population variability (thin lines) is compared with replicate variability (thick lines) for the population model (1) with $x\,(0) = 1$ (as in figures 1 and 3). Variance is shown in (a) and variance-to-mean ratio in (b). Note the logarithmic scale on the $y$-axis. In this figure, 10 randomly chosen replicates are analysed for replicate variability, in contrast to only one replicate in figure 3. Only part of the graph for the population variance-to-mean ratio is shown in (b); for the full plot compare with figure 3(b).

This behaviour can be contrasted with a model in which there is no nonlinear growth. For the Ornstein–Uhlenbeck process characterised by weak correlation ($\lambda = 0.9$; figures 5(a) and 5(b)), the ensemble variance and the replicate variance are both constant in time and approximately equal, suggesting that short-term and long-term predictability is the same in this case. This corresponds to the traditional approach used in modelling population dynamics, where each observation in each replicate is assumed to be uncorrelated with other observations. When strong temporal correlations are included ($\lambda = 1.0$, a Wiener process, figures 5(c) and 5(d)), the ensemble variance is larger than the replicate variance. The former increases linearly in time, whereas the latter stays constant. The population mean is close to 0 in both cases. In the case of the Wiener process, the short-term predictability is much better than the long-term predictability, due to strong temporal correlations.

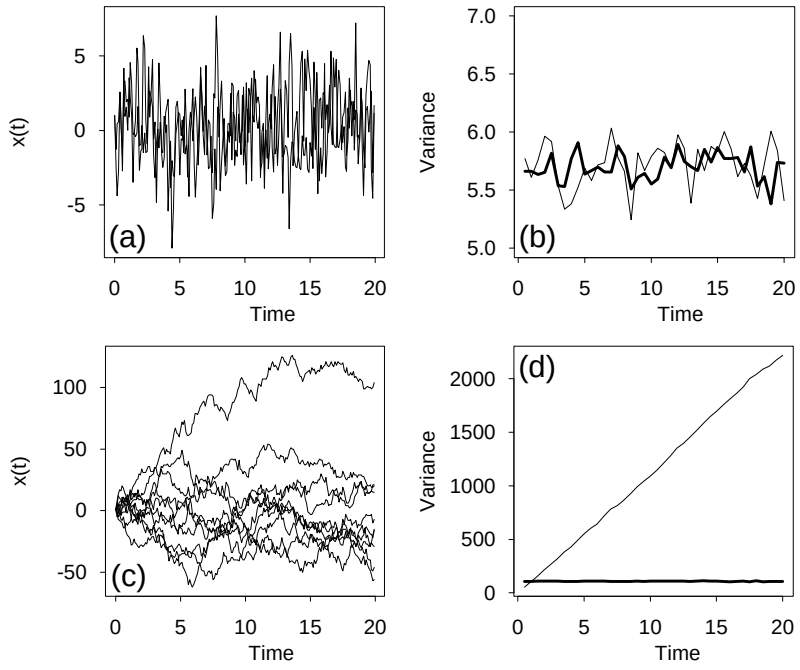


Fig. 5. Examples of trajectories for the Ornstein–Uhlenbeck process with $\lambda = 0.9$ (a) and for the Wiener process, $\lambda = 1.0$ (b). Only two trajectories are shown in (a) and ten trajectories in (b), for clarity of presentation. Population variances (thin lines) are compared with replicate variances (thick lines) for the OU process (c) and for the Wiener process in (d). Note that in contrast to figure 2, the $y$-axis is not logarithmic here.

## 4. Discussion

We have shown that for a model that describes exponential growth, the variability along a single replicate is much smaller than between realisations, and the difference increases as the population becomes larger.

The striking difference in the asymptotic behaviour of the ensemble and replicate variances can be understood by identifying the main sources of variability in both cases. When the process $x(t)$ is simulated between $t$ and $t' > t$ ($t' - t$ small), the uncertainty in the prediction is associated with the Poisson process for which the variance equals the mean. Since for the short-term prediction — based on the current value of $x_k(t)$ — the Poisson error is the dominating factor, the replicate variance follows the mean value (as in figure 3(b)).

On the other hand, figures 1(c) and 2(c) suggest that the ensemble variability is mostly determined in the first phase of the dynamics, when the populations are small. Because the mechanism for the generation of the initial variability is also Poissonian, the initial variance is related to the mean population size at $t = 0$, $x(0)$. With time growing, the within-replicate variability becomes proportionally smaller (figures 1(b) and 1(c)), and so the trajectories become effectively smoother. At the same time, the individual replicates grow exponentially, and therefore the difference between two randomly picked realisations also grows exponentially like $\exp(bt)$. As a result, the ensemble variance follows the square of the mean $x^2(t)$, rather than the mean. Thus, the different behaviour of the ensemble and replicate behaviour is a result of a small initial sample size (contrast figures 1 and 2) and a nonlinear (exponential) growth. This analysis allows us to identify the basic mechanisms responsible for such a difference: small initial population size and nonlinear (exponential) departures of trajectories. The results reported here are consistent with analytical solutions given in [8] and in the Appendix A below.

The results presented here do not contradict any asymptotic results obtained in the limit of large population sizes. Any results, like those in [10], assume that $x(t) \gg 1$ for *all* times. In our case, the initial value of $x(t)$ is small.

The population model we used in this paper is very simple, assuming no limitation to the growth of a population. However, exponential function forms a basic building block of many population models. More complicated models, including those studied in [4, 6] can also be simulated and analysed using our method. We suggest that our analysis has profound consequences for consideration of appropriate statistical methods in model fitting as well as for the design and interpretation of ecological and epidemiological experiments [4].

The work was funded by DEFRA (UK). The paper originates from my collaboration with D.J. Bailey, G.J. Gibson, C.A. Gilligan, P.F. Góra, E. Gudowska-Nowak and W. Otten and I am very grateful to them for all suggestions. I am also very grateful to Professor Andrzej Fuliński for his scientific and personal support. Working under His caring guidance was the highlight of my scientific career.

## Appendix A

*Analytical expressions for the 'ensemble' variability*

The results presented here are obtained in [2, 8]. The probability that $x(t)$ equals to $x$, given the initial condition $x(0) = x_0$ is given by

$$P(x(t) = x) = \binom{x-1}{x_0-1} \left(e^{-bt}\right)^{x_0} \left(1 - e^{-bt}\right)^{x-x_0}. \tag{A.1}$$

Mean value of $x(t)$ follows the following equation

$$\mathrm{E}[x(t)] = x_0\, e^{bt}. \tag{A.2}$$

Variance is given by

$$\mathrm{Var}[x(t)] = x_0\, e^{2bt}\left(1 - e^{-bt}\right) \to x_0\, e^{2bt}, \tag{A.3}$$

when $t \to \infty$. As a result,

$$\frac{\mathrm{Var}[x(t)]}{\mathrm{E}[x(t)]} = e^{bt} - 1 \to \frac{1}{x_0}\,\mathrm{E}[x(t)], \tag{A.4}$$

and

$$\frac{\sqrt{\mathrm{Var}[x(t)]}}{\mathrm{E}[x(t)]} = (x_0)^{-1/2}\,\sqrt{1 - e^{-bt}} \to (x_0)^{-1/2}, \tag{A.5}$$

all in the long-time limit.

## REFERENCES

[1] A. Kleczkowski, *Acta Phys. Pol. B* **29**, 1717 (1998).

[2] G. Grimmet, D. Stirzaker, *Probability and Random Processes*, Oxford University Press, 2001.

[3] R. Streater, *J. Math. Phys.* **41**, 3556 (2000).

[4] A. Kleczkowski, D.J. Bailey, C.A. Gilligan, *Proc. Royal Soc.* **B263**, 777 (1996).

[5] I. Epstein, *Nature* **374**, 321 (1995).

[6] C.A. Gilligan, A. Kleczkowski, *Phil. Trans. Royal Soc.* **B352**, 591 (1997).

[7] C. Gardiner, *Handbook of Stochastic Methods*, Springer Verlag, Berlin, Heidelberg, New York 1985.

[8] G. Grimmet, D. Stirzaker, *One Thousand Exercises in Probability*, Oxford University Press, 2001.

[9] V. Barnett, *Comparative Statistical Inference*, Wiley, 1999.

[10] J.P. Aparicio, H.G. Solari, *Phys. Rev. Lett.* **86**, 4183 (2001).