The effect of sampling rate on observed statistics in a correlated random walk

G. Rosser, A. G. Fletcher, P. K. Maini and R. E. Baker

Centre for Mathematical Biology, Mathematical Institute, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, UK

Tracking the movement of individual cells or animals can provide important information about their motile behaviour, with key examples including migrating birds, foraging mammals and bacterial chemotaxis. In many experimental protocols, observations are recorded with a fixed sampling interval and the continuous underlying motion is approximated as a series of discrete steps. The size of the sampling interval significantly affects the tracking measurements, the statistics computed from observed trajectories, and the inferences drawn. Despite the widespread use of tracking data to investigate motile behaviour, many open questions remain about these effects. We use a correlated random walk model to study the variation with sampling interval of two key quantities of interest: apparent speed and angle change. Two variants of the model are considered, in which reorientations occur instantaneously and with a stationary pause, respectively. We employ stochastic simulations to study the effect of sampling on the distributions of apparent speeds and angle changes, and present novel mathematical analysis in the case of rapid sampling. Our investigation elucidates the complex nature of sampling effects for sampling intervals ranging over many orders of magnitude. Results show that inclusion of a stationary phase significantly alters the observed distributions of both quantities.

1. Introduction

Tracking is a widely used experimental method to probe the movement of living organisms over a huge range of spatial scales, from bacteria swimming over distances of tens of micrometres [1], to birds migrating thousands of kilometres [2]. While the details of the experimental protocol vary between different studies, in the majority of cases the data obtained are similar, namely an ordered list of position vectors (observations) sampled at discrete time points that approximates the continuous-time underlying motion. The time intervals between observations may have a profound impact on the movement patterns observed and the conclusions drawn [3–5]. It is therefore crucial to characterize these effects. We restrict our attention to cases where the sampling interval between successive observations is constant, as is the case in many tracking systems, including global positioning systems [6] and microscopy [1].

Tracking generates large quantities of information about the motion of individuals and, where sufficiently many tracks are available, populations. These data may be interrogated in order to extract statistics and test hypotheses specific to the organism being studied. Examples include the homing behaviour of rockfish [7] and oscillations in the dive depths of sharks [8]. To ensure the wide applicability of the present work, we consider the effect of the sampling rate on two quantities that are of fundamental interest in many such studies: the apparent speed of an individual between two consecutive observations and the apparent angle change (AAC) between three consecutive observations. These quantities have been used to elucidate the chemotactic response of the bacterium *Rhodobacter sphaeroides* [9] and the movement of beetles close to the boundaries of habitats [10]. Our work also applies to apparent displacement, since for a given sampling interval this quantity is linearly proportional to the apparent speed.

Any theoretical study of the tracking process requires the specification of a model of motion. Here, we consider individuals undergoing a correlated

random walk (CRW) [11]. Various species move in a manner well described by a CRW, in the sense that their motility takes the form of approximately straight line movements interspersed by stochastic reorientations. This includes the 'run-and-tumble' pattern of motility exhibited by many species of planktonic bacteria [12,13], ovipositing butterflies [14], foraging bumblebees [15] and the clonal growth of the plant species *Solidago altissima* [16]. Correlation is introduced because turning angles are specified relative to the previous direction of motion. Turning angles are also commonly assumed to be independent and identically distributed (iid), hence the correlation exists between successive positions, rather than successive angle changes. In many studies, it is assumed that reorientations occur instantaneously. Here, we consider both this model and a variant, proposed by Othmer *et al.* [13], in which reorientations are concomitant with a stationary 'resting' phase. This is a good description of the motion of bacteria such as *R. sphaeroides* that stop swimming to reorient [9], and elk that pause at foraging sites to feed [17].

The effect of sampling rate on observed tracks is a complex problem that has, to date, received relatively little research attention. An important study performed by Kareiva & Shigesada [14] considers a CRW with variable speed, deriving an exact expression for the mean squared displacement (MSD) after a specified number of steps, in terms of the mean cosine of angle changes, mean step length and mean squared step length. This work was extended by McCulloch & Cain [18], who derive an approximate expression for the mean displacement. Bovet & Benhamou [15] study a CRW with fixed step lengths between observations. In order to make tracks compatible with this description a spatial rediscretization is required, leading to variable time intervals between observation points. The authors obtain an empirical expression for the variation of angular standard deviation with step length.

While the above studies consider the effect of number of steps or step length (spatial rediscretization), in the present work we are concerned with the effect of sampling rate (temporal rediscretization). In the first study to consider this problem, Hill [19] rediscretizes tracks with a variable sampling interval and makes the ad hoc assumption that the mean and standard deviation of apparent speeds vary linearly with the size of this interval. More recently, Codling & Hill [12] have analysed the effect of the sampling rate on a CRW, repeating the analysis of Bovet and Benhamou with a temporal rediscretization of tracks. The authors use simulated tracks to find empirical expressions for the variation of the standard deviation of AACs, and the mean and standard deviation of apparent speeds. They demonstrate that Hill and Häder's assumption of a linear relationship between these properties and the sampling interval is only valid for a limited range of sampling intervals.

1.1. Aims and outline

In this study, we address several of the open problems in this field. In §3, we obtain the stationary distributions of the apparent speeds and AACs for the run-only model in the limit of large sampling intervals. In §4, we extend the work of Codling & Hill [12], by computing numerically the full distribution of these quantities, rather than the first two moments, thus fully characterizing their complex dependence on the sampling interval. We also consider a modified model, in which reorientations are accompanied by a stationary phase.

Table 1. Glossary of notation.

parameter	description
τ	sampling time interval
τ_R	mean time between reorientation events
τ_S	mean duration of a stationary reorientation event
c_{const}	underlying constant run speed
$\bar{c}$	mean relative apparent speed (RAS)
σ_c	standard deviation of RASs
σ_δ	underlying angular deviation
σ_θ	apparent angular deviation
ρ	mean cosine of apparent angle changes (AACs)

This has, to our knowledge, not previously been investigated. In §5, we present a novel analytic approach to describe the apparent speed and AAC distributions for the run-only model in the limit of rapid sampling.

2. Models and methods

A summary of all mathematical notation used in this work is given in table 1. In this study, we investigate the effect of the fixed sampling interval, denoted by τ, on the observations made from a tracking experiment of an individual moving with a constant speed, c_{const}, in a two-dimensional unbiased CRW. Reorientation events occur as a Poisson process with constant mean inter-arrival time τ_R. Following Codling & Hill [12], we consider the non-dimensional ratio τ/τ_R, which represents the mean number of reorientations in a sampling interval. We examine two variants of the model, proposed by Othmer *et al.* [13]: (i) the 'run-only' model, in which reorientation events occur instantaneously and (ii) the 'run-and-stop' model, in which reorientations take a finite time, during which the individual is stationary. In the run-and-stop model, the individual also leaves the stationary phase as a Poisson process, with mean inter-arrival time τ_S. Figure 1 shows a simulated run-only track (see §2.1 for details). Between consecutive reorientation events, an individual covers a random straight-line distance L, which is distributed exponentially with mean $c_{const}\tau_R$. During a reorientation event, an individual turns through an angle $\Phi \in [-\pi, \pi]$, a random variable with a probability density function (pdf) given by $f_\Phi(\phi)$. We assume that the angle changes are iid and independent of the running distance, L. In the run-and-stop model, we also assume that angle changes occur independently of the duration of the reorientation phase.

In §§4 and 5, we assume the underlying angle changes Φ are drawn from the Von Mises (VM) distribution, a commonly used distribution in directional statistics [12,14,15]. It is a good approximation for the normal distribution on a circle and mathematically more convenient, as the pdf does not involve an infinite summation [20]. We choose the VM distribution with pdf

$$f_\Phi(\phi; \kappa) = \frac{e^{\kappa \cos(\phi)}}{2\pi I_0(\kappa)}, \quad (2.1)$$

where I_0 denotes the modified Bessel function of the first kind with order zero and the parameter $\kappa \geq 0$ controls the
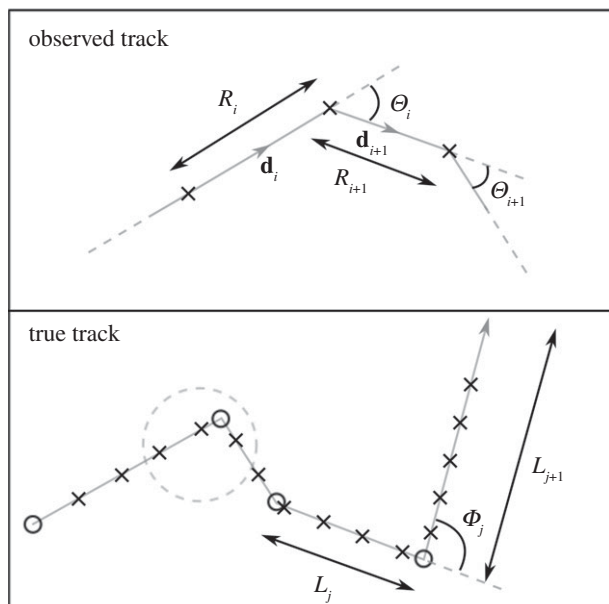
Figure 1. The effect of sampling on apparent motion. The lower track is simulated using the run-only model. Circles indicate reorientation events, crosses indicate sampling points and two run lengths and a turning angle are shown. The upper panel shows the circled portion in greater detail. This is the observed track, with two apparent displacements and two AACs marked.

'peakedness' of the distribution. This distribution is symmetric about zero, hence has zero mean, since our motion model is an unbiased CRW. When dealing with circular statistics, alternative summary statistics are needed from those used in linear statistics. We shall use the angular deviation [20], given by

$$\sigma_\delta = \sqrt{2(1-\rho)}, \qquad (2.2)$$

where ρ denotes the mean cosine of the underlying angle change, Φ. For the VM distribution (2.1), we have [20]

$$\rho = \frac{I_1(\kappa)}{I_0(\kappa)}. \qquad (2.3)$$

We may also use equation (2.2) to compute the angular deviation of the AACs, σ_θ, in which case ρ denotes the apparent mean cosine. For consistency, we therefore specify the angular deviation of the underlying VM distribution, σ_δ, which we convert to the VM parameter κ using equations (2.2) and (2.3). This equation has a unique solution as $I_1(\kappa)/I_0(\kappa)$ is monotonic in κ; we compute this using a numerical optimization routine based on the trust-region algorithm [21].

The sampling process is modelled as a uniform temporal discretization of the underlying track, with consecutive observations separated by the sampling interval τ. Since reorientations occur stochastically in continuous time, observations do not in general coincide with reorientation events. The observed track takes the form of a series of transition vectors, denoted $\mathbf{d}_i$, where $i \in \{1, 2, \ldots\}$ gives the ordering of the transitions. The quantities of interest are the relative apparent speed (RAS) between consecutive observations, $R_i/c_{\mathrm{const}} = ||\mathbf{d}_i||/c_{\mathrm{const}}\tau$, and the AAC between three consecutive observations, $\Theta_i = \arccos(\mathbf{d}_i \cdot \mathbf{d}_{i+1}/||\mathbf{d}_i||\ ||\mathbf{d}_{i+1}||)$ (figure 1). The RAS is a non-dimensional measure of the proportion of the true speed obtained by the observed individual across a sampling interval; use of this quantity ensures that our results

are independent of c_{const}. We record all observed quantities in order to approximate the pdf by computing a histogram. We also compute the mean RAS, $\bar{c}$, the standard deviation of RASs, σ_c, and the angular deviation of AACs, σ_θ.

When computing summary statistics for the run-and-stop CRW, we discard AACs and speeds arising from static intervals in which the simulated individual has zero displacement. Static intervals in tracks are readily discernable, and provide a direct means to calculate the underlying reorientation angle. Where the tracking data permit this type of direct analysis, for example, when reorientations occur at landing sites [14], other analysis methods are more appropriate than computing all apparent quantities.

For simplicity, we neglect the effects of measurement noise; while the effects of noise and sampling frequency are related [4], many open questions remain about the role of sampling frequency alone. We discuss this in further detail in §6.

2.1. Stochastic simulation algorithm

Throughout this study, we use the following stochastic simulation algorithm to generate realizations of the unbiased two-dimensional CRW. In the run-only model, waiting times between reorientation events are drawn from the exponential distribution with mean time τ_R. In the run-and-stop model, the duration of each stationary reorientation event is also exponentially distributed with mean τ_S. Reorientation events are simulated by drawing an angle change from the VM distribution with concentration parameter κ using the algorithm of Best & Fisher [22]. Between reorientations, the individual moves in the chosen direction with constant speed, $c_{\mathrm{const}} = 1\,\mathrm{m\,s}^{-1}$. The stochastic trajectory of a single individual is simulated in this way until the total time exceeds a predefined value, at which point the trajectory is ceased. The position of each simulated individual is finally calculated at regularly spaced times, separated by the sampling interval τ.

All simulations are implemented in Matlab. The code used to simulate tracks is available in the electronic supplementary material, in addition to a movie illustrating the simulated tracks.

3. Stationary sampling distributions

When a run-only CRW is sampled at low frequency, many reorientation events occur between consecutive observations and hence persistence in the observed motion is lost. The observed process is diffusive, provided that the mean cosine of the underlying angle change distribution satisfies $\rho < 1$ [11,23]. We now describe the asymptotic behaviour of the CRW in the diffusive limit. Consider an individual undergoing two-dimensional Brownian motion, observed at regular time intervals, τ. Between each observation, the displacements in each dimension, d_x and d_y, are iid and normally distributed, with zero mean and variance equal to $2D\tau$, where D is the macroscopic diffusion coefficient. Othmer *et al.* [13] have previously shown that, for a constant-speed two-dimensional run-only CRW, $D = c_{\mathrm{const}}^2 \tau_R/(2(1-\rho))$. The RAS between consecutive sample points, $\sqrt{d_x^2 + d_y^2}/(c_{\mathrm{const}}\tau)$, is therefore distributed according to the scaled χ distribution with two degrees of freedom and scale parameter $\sqrt{\tau/(\tau_R(1-\rho))}$. The
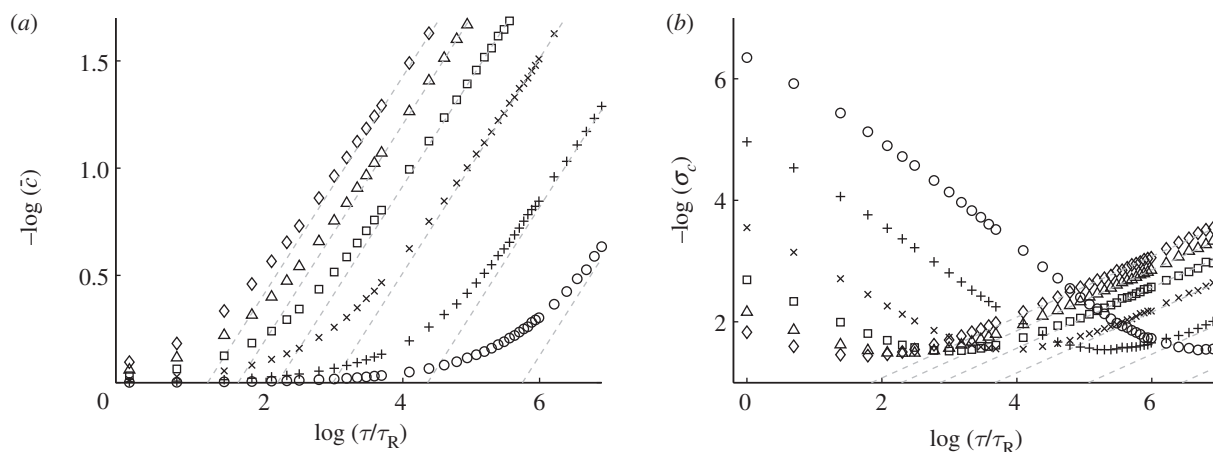
Figure 2. (a) Log–log plot of the mean RAS against sampling interval. (b) Log–log plot of the standard deviation of RASs against sampling interval. The data are computed for a run-only CRW with $\sigma_\delta = 0.1$ (open circles), $\sigma_\delta = 0.2$ (plus symbols), $\sigma_\delta = 0.4$ (cross symbols), $\sigma_\delta = 0.6$ (open squares), $\sigma_\delta = 0.8$ (open triangles) and $\sigma_\delta = 1.0$ (open diamonds). Dashed lines show the asymptotic behaviour, computed using (3.1) and (3.2).

mean of this distribution is given by

$$\bar{c} = \sqrt{\frac{2}{\tau/(\tau_R(1-\rho))}} \Gamma\!\left(\frac{3}{2}\right), \tag{3.1}$$

and its standard deviation is given by

$$\sigma_c = \sqrt{\frac{2}{\tau/(\tau_R(1-\rho))}\left(1 - \Gamma\!\left(\frac{3}{2}\right)^2\right)}. \tag{3.2}$$

Expressions (3.1) and (3.2) are plotted in figure 2 overlaid with stochastic simulation results for a range of values of σ_δ. As σ_δ increases, so does the time scale of relaxation to the diffusion limit. This result is intuitive; trajectories with high angular deviation lose persistence rapidly, as reorientations are often drastic, hence the time scale at which they appear diffusive is short compared to a trajectory with low angular deviation. Regardless of the value of σ_δ, the standard deviation of RASs initially increases with sampling interval, before eventually decreasing at sampling intervals near to the diffusive limit. We return to this observation in §4.1.

For a diffusing individual, the AACs between consecutive pairs of observations are distributed uniformly on a circle [23]. The wrapped uniform distribution has a mean cosine of zero, therefore from equation (2.2) the apparent angular deviation is $\sqrt{2}$. This result is plotted in figure 3. As for the mean RAS, the greater the true angular deviation, the more quickly the observed process relaxes to the diffusion limit.

The results presented above may be extended to include a stationary reorientation phase in a straightforward manner, by noting that the mean proportion of time spent in a running state is given by $\tau_R/(\tau_R + \tau_S)$. Multiplying the macroscopic diffusion coefficient D by this quantity gives the effective diffusion coefficient for a run-and-stop CRW.

4. Simulation study of dynamic sampling distributions

For many applications, the apparent motile behaviour of individuals at shorter sampling intervals is of greater practical interest than in the diffusive limit. Having derived results for the asymptotic properties of the run-only CRW when τ/τ_R is large, we now consider the effect of sampling on time scales
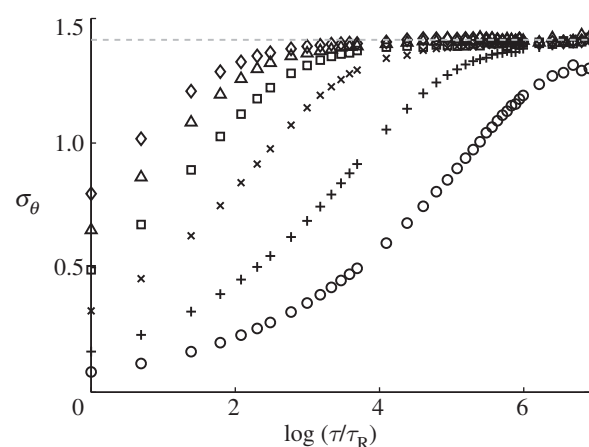
Figure 3. A plot of the apparent angular deviation against log sampling interval for a run-only CRW. Symbols indicate the values of σ_δ, as in figure 2. The dashed line indicates the asymptotic value of $\sigma_\theta = \sqrt{2}$.

corresponding to $\tau/\tau_R \geq 1$. Here, persistence in the motion is perceivable and the observed process is not purely diffusive. We shall use stochastic simulations of the run-only and run-and-stop CRW models to compute the pdf of RASs and AACs for a range of values of τ/τ_R.

4.1. Run-only model

The distribution of RASs for different sampling intervals is shown in figure 4a for two different values of σ_δ. There is a greater spread in the case $\sigma_\delta = 0.4$ compared with the case $\sigma_\delta = 0.1$, because each reorientation event tends to result in a larger change of direction. As τ/τ_R is increased, the RAS distribution becomes broader because each sampling interval has a greater probability of containing one or more reorientation events. When a reorientation occurs in an interval, the RAS is reduced.

The results in figure 4a are in agreement with the findings of Codling & Hill [12]. As we increase τ/τ_R, the distribution of RASs is skewed towards lower speeds. This decreases the mean RAS. We further investigate this phenomenon in figure 5, in which we show the variation of the first percentile RAS with τ/τ_R. When the underlying angular deviation is high, the first percentile RAS is less than 0.5 for even the highest sampling rate considered. Conversely, when $\sigma_\delta = 0.1$, the first
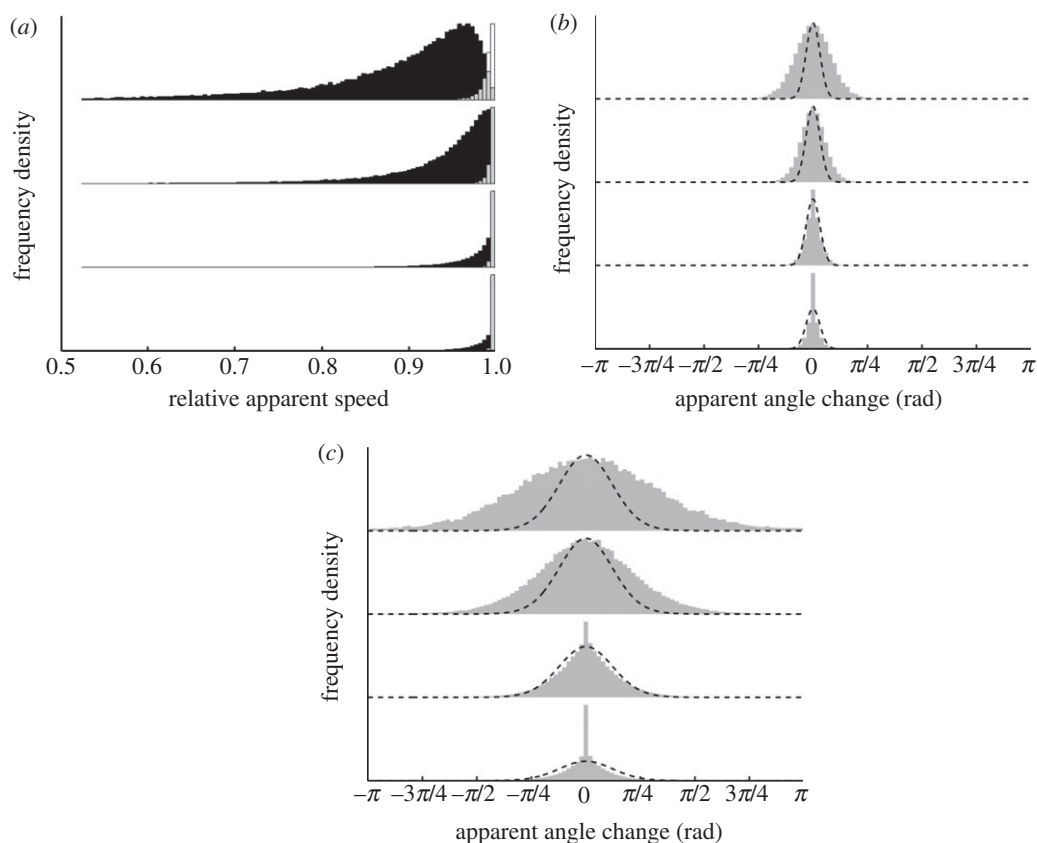
Figure 4. RAS and AAC distributions for different sampling intervals, computed with two different values of σ_δ. Stacked distributions correspond to different values of τ/τ_R: (bottom to top) 1, 2, 5, 10. The y-axis scale varies between stacked plots. (a) RAS distribution. Black bars correspond to $\sigma_\delta = 0.4$, white bars correspond to $\sigma_\delta = 0.1$, grey bars indicate overlap. AAC distribution for (b) $\sigma_\delta = 0.1$ and (c) $\sigma_\delta = 0.4$. The dashed line indicates the true underlying angle change distribution used in simulations.
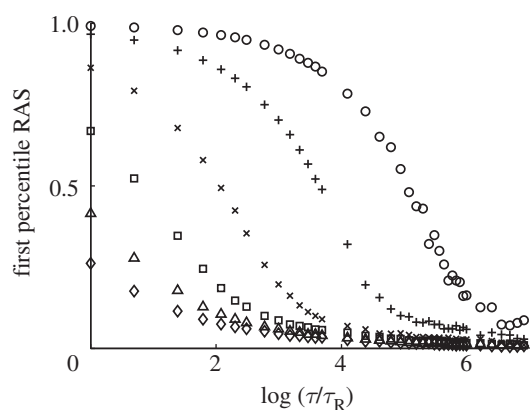
Figure 5. Plot showing the variation in the first percentile RAS with sampling interval. The data are computed for a run-only CRW with $\sigma_\delta = 0.1$ (open circle), $\sigma_\delta = 0.2$ (plus symbol), $\sigma_\delta = 0.4$ (cross symbol), $\sigma_\delta = 0.6$ (open square), $\sigma_\delta = 0.8$ (open triangle) and $\sigma_\delta = 1.0$ (open diamond).

percentile RAS remains close to one until the sampling interval is relatively large. This is a key example of the use of alternative summary statistics to demonstrate the effects of sampling rate. The insight gained from considering the full distribution of RASs now allows us to rationalize the variation of the standard deviation of RASs, plotted in figure 2b. The RASs are tightly clustered about the true value of c for small τ/τ_R, becoming increasingly skewed towards lower RASs as τ/τ_R increases, leading to the initial increase in σ_c with τ/τ_R. As the sampling interval becomes large, RASs are observed in the vicinity of zero with increasing frequency, which reduces σ_c, leading to the observed asymptotic behaviour.

As we sample more frequently, the distribution of RASs converges to the true underlying distribution, as individuals undergo fewer reorientations between sampling intervals. For the constant-speed process considered here, the distribution of RASs therefore tends towards a delta function at the true underlying speed as we decrease τ/τ_R. This result is not restricted to a constant-speed process; when there is an underlying distribution of speeds, the observed speed distribution converges to the true underlying distribution (data not shown).

Figure 4b,c show the distribution of AACs for several values of τ/τ_R and σ_δ. The VM distribution used for reorientations in the underlying CRW is also shown. Note that this distribution is constant in each figure; the change in height is because of a rescaling of the y-axis. The broadening of the distribution of AACs results in an increase in σ_θ. In the rapid sampling limit, this distribution tends towards a delta function at the origin, because most sampling intervals contain no reorientations and hence measure zero angle change. Such intervals are described as 'artificial zero turns' by Codling & Hill [12]. In contrast with RASs, where the true underlying distribution is obtained in the limit $\tau/\tau_R \to 0$, at no sampling interval does the distribution of AACs match the true underlying angle change distribution. This is because, in making observations at regular intervals, we sample both across and between reorientations.

4.2. Run-and-stop model

We now perform a similar analysis for the run-and-stop model, in which the additional parameter τ_S specifies the
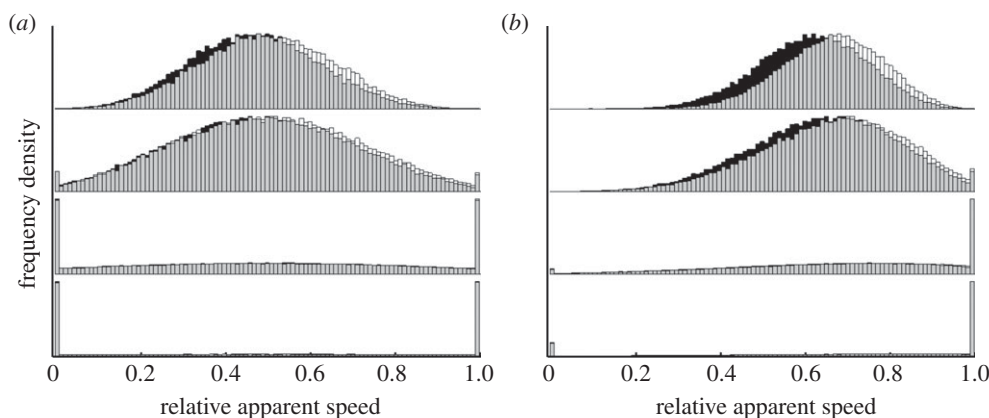
Figure 6. Apparent speed distributions for different sampling intervals, computed with (*a*) $\tau_S = 1$ s and (*b*) $\tau_S = 0.5$ s. Stacked distributions correspond to different values of τ/τ_R: (bottom to top) 1, 2, 5, 10. The *y*-axis scale varies between stacked plots. Black bars correspond to $\sigma_\delta = 0.4$, white bars correspond to $\sigma_\delta = 0.1$, grey bars indicate overlap.
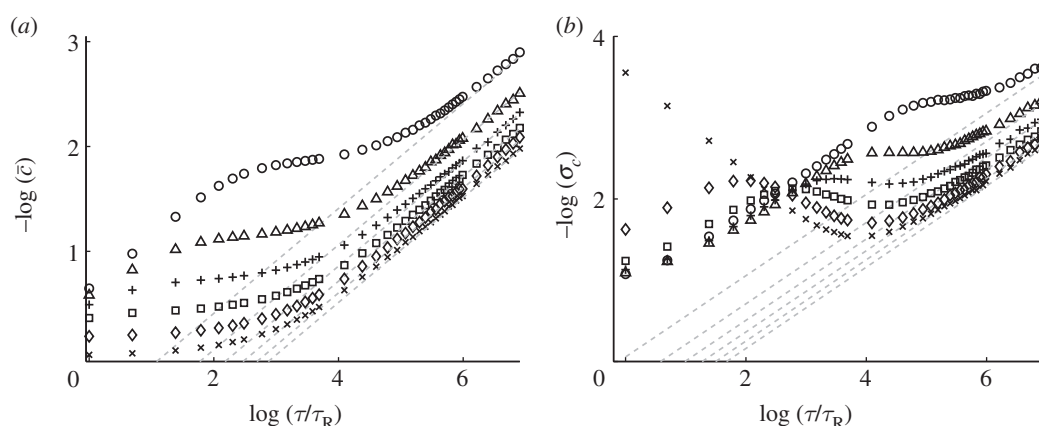
Figure 7. Plots showing the variation of the mean RAS (*a*) and standard deviation of RASs (*b*) with τ/τ_R for a run-and-stop CRW with $\sigma_\delta = 0.4$ and $\tau_S = 0$ s (cross symbols, equivalent to a run-only process), $\tau_S = 0.2$ s (open diamonds), $\tau_S = 0.5$ s (open squares), $\tau_S = 1$ s (plus symbols), $\tau_S = 2$ s (open triangles) and $\tau_S = 5$ s (open circles). Dashed lines show the asymptotic behaviour, computed using (3.1) and (3.2).

mean duration of a stopping phase. The distributions of RASs for two values of σ_δ and τ_S are shown in figure 6. Compared with the run-only model (figure 4), two peaks are visible at $c = 0$ and 1, corresponding to individuals that are stationary for the duration of an interval and to individuals that undergo no reorientations in an interval, respectively. Both decay as τ/τ_R increases. The density at intermediate RASs spans the full range of permissible values, in contrast to the run-only case, in which the range of observed values is more limited. When $\tau_S = \tau_R$, the densities in the two outer peaks are equal, as individuals populate both the running and reorientating states equally, on average. Reducing τ_S causes the distribution to be more negatively skewed for all values of τ/τ_R considered, as individuals spend more time running than reorienting.

In figure 7, we plot the mean and standard deviation of RASs, $\bar{c}$ and σ_c, for different values of τ_S. As discussed earlier, we exclude static intervals in the calculation of these quantities. Comparison of figures 2*a* and 7*a* indicates that the time scale on which the process appears diffusive is longer in the run-and-stop CRW than the run-only CRW. In all cases, the mean RAS decreases monotonically for all sampling intervals considered, with the initial rate of decrease becoming more marked for larger τ_S.

We find that σ_c displays a non-monotonic dependence on the sampling rate for $\tau_S = 0.2$, 0.5 and 1 s. For small sampling

intervals, the RAS pdf is dominated by the two peaks at 0 and 1, with resulting large standard deviation. As the sampling interval increases, the density between these peaks initially increases, hence σ_c decreases. When $\log(\tau/\tau_R) \approx 2$, the outer peaks are no longer apparent, the density in the central region continues to broaden with sampling interval, and hence σ_c increases with τ/τ_R. At yet longer sampling intervals corresponding to $\log(\tau/\tau_R) \gtrsim 4$, RASs are increasingly observed in the vicinity of 0, therefore σ_c again decreases with τ/τ_R, as predicted in §3 and discussed in the case of the run-only model.

The distribution of AACs is shown for $\sigma_\delta = 0.4$ and $\tau_S = 1$ s in figure 8*a*. The distributions shown do not include angle changes arising from static intervals. Comparing with figure 4, we see that the inclusion of a stopping phase causes the distribution of observed angle changes to become more concentrated about $\theta = 0$. The shape of the distribution is, however, qualitatively similar in both cases. Figure 8*b* shows the dependence of σ_θ on sampling frequency. As for the run-only CRW, an asymptote is reached at $\sigma_\theta = \sqrt{2}$, though the time scale of relaxation to this diffusive limit increases with τ_S. Furthermore, as τ_S increases, the angular deviation decreases for a given value of τ/τ_R. A larger value of τ_S means that, on average, fewer reorientations occur over a sample interval. As a result, the distribution of AACs has greater density at low values, and the apparent angular deviation is lower.
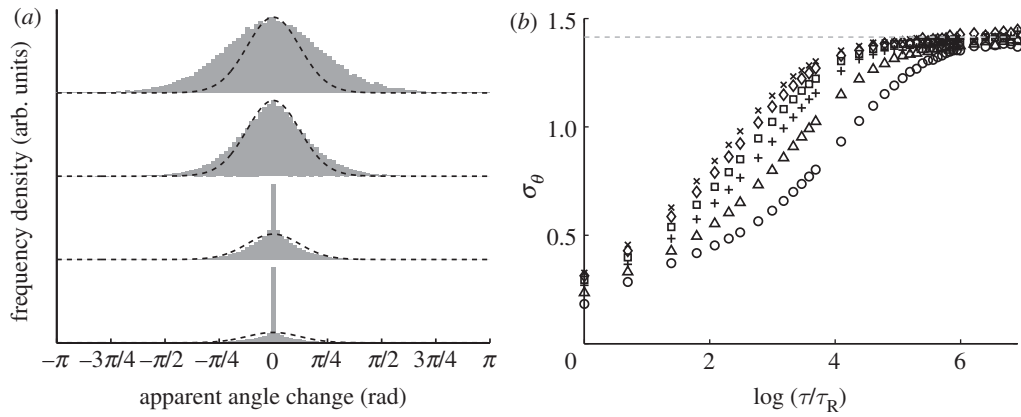
Figure 8. The variation of AACs with sampling interval for a run-and-stop CRW. (a) AAC distributions for different sampling intervals, with $\sigma_\delta = 0.4$ and $\tau_S = 1$ s. Dashed line indicates the true underlying angle change distribution. Stacked distributions correspond to different values of τ/τ_R: (bottom to top) 1, 2, 5, 10. Note that the y-axis scale varies between plots. (b) The variation of apparent angular deviation, σ_θ, with τ/τ_R for $\sigma_\delta = 0.4$, $\tau_R = 1$ s and $\tau_S = 0$ s (cross symbols, equivalent to a run-only process), $\tau_S = 0.2$ s (open diamonds), $\tau_S = 0.5$ s (open squares), $\tau_S = 1$ s (plus symbols), $\tau_S = 2$ s (open triangles) and $\tau_S = 5$ s (open circles). The dashed line indicates the asymptotic value of $\sigma_\theta = \sqrt{2}$.

# 5. Analytic study of dynamic sampling distributions

While stochastic simulations are useful for assessing the effect of sampling rate on the observed CRW, they are time-consuming, as many realizations are required to obtain sufficiently smooth distributions. Furthermore, a new simulation is required for each new set of parameters, making it hard to draw general conclusions about the effect of varying the reorientation angle pdf. We therefore seek an analytic description of the underlying CRW, for application to the sampling rate problem. As before, we consider an unbiased, run-only, constant-speed, two-dimensional CRW. We include a general underlying pdf for reorientations, $f_\Phi(\phi)$. In order to make analytic progress, we assume that the relative sampling interval, τ/τ_R, is sufficiently small that the probability of two or more reorientations occurring between consecutive sample points is negligible. Using the fact that reorientations in the underlying CRW occur as a Poisson process, this probability is given by

$$\mathbb{P}(N \geq 2; \tau, \tau_R) = 1 - e^{-\tau/\tau_R}(1 + \tau/\tau_R), \qquad (5.1)$$

where N denotes the number of reorientations in a time interval of duration τ. We also assume that no two consecutive sampling intervals contain reorientation events, an occurrence with probability given by

$$\mathbb{P}(N > 0 \text{ in two consecutive intervals}; \tau, \tau_R) = (1 - e^{-\tau/\tau_R})^2. \qquad (5.2)$$

Expression (5.2) is always greater than (5.1). For example, choosing $\tau/\tau_R = 0.05$, (5.1) evaluates to 1.2×10^{-3}, and (5.2) gives a value of 2.4×10^{-3}, so one of the two assumptions is broken once every 280 intervals, on average. Discounting these events, all intervals are assumed to contain at most one reorientation event, with no consecutive events. A key example of an experimental protocol that is compatible with these assumptions is bacterial motility, which is commonly probed using video microscopy to perform tracking [1]; in these studies, the frame capture rate of the digital microscope camera is typically significantly higher than the rate at which bacteria reorient [1,24,25].
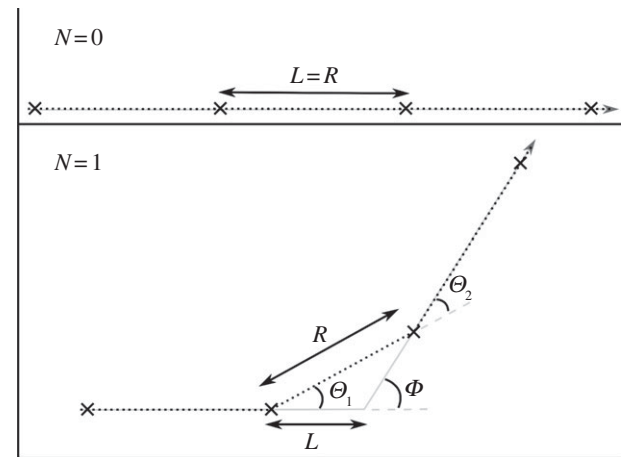
Figure 9. Illustration of portions of a CRW trajectory containing zero and one reorientation phases. Grey lines show the true path, dotted black lines show the observed path, and the crosses indicate the sampling points. See main text for further explanation.

Figure 9 shows two portions of a simulated trajectory corresponding to $N = 0$ and 1. The former case always returns AACs equal to 0 and step lengths equal to $c_{const}\tau$. In the case of a single reorientation event, the true underlying trajectory consists of two straight line sections separated by an instantaneous reorientation of angle Φ, given relative to the previous direction of travel. In contrast, the observed path appears to contain two reorientation events, with angles Θ_1 and $\Theta_2 = \Phi - \Theta_1$. The statistical properties of the unbiased CRW are time-reversal invariant, meaning that the distributions of Θ_1 and Θ_2 must be identical. We therefore drop the subscript in all further notation. The apparent displacement is denoted by R. As before, the difference between the apparent and true paths is because of the discrete sampling of the continuous underlying trajectory. Let L denote the distance between the last sampling point and the reorientation event. For a constant-speed process, the distance travelled by an individual over the course of a sampling interval is given by $c_{const}\tau$. The distance between the reorientation event and the next sampling point is therefore $c_{const}\tau - L$. We reiterate that the CRW model used in the present study assumes Φ is independent of L.
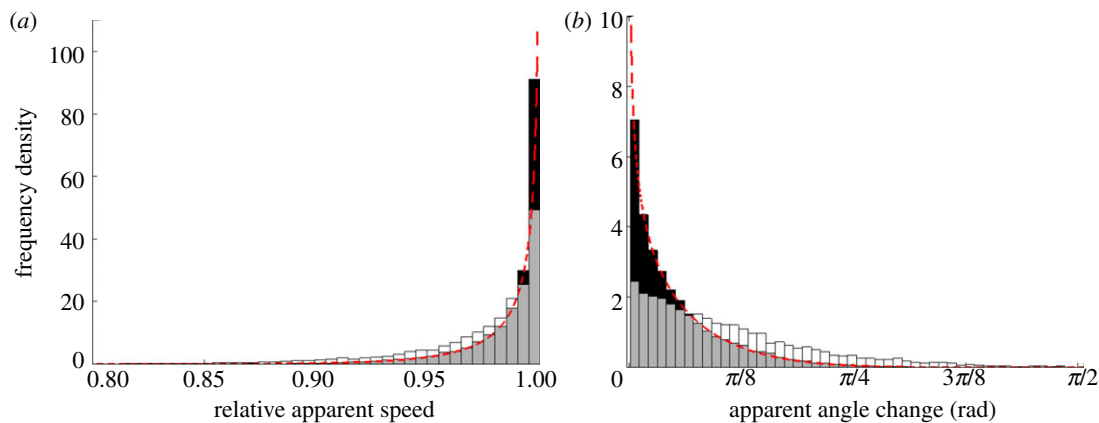
Figure 10. Distribution of RASs (*a*) and absolute AACs (*b*), conditional upon a single reorientation event occurring between consecutive sampling intervals, computed by numerical integration of equation (5.9) for a CRW with $\sigma_\delta = 0.4$. In both cases the results from a stochastic simulation are shown after subtracting zero-stop events (see main text for explanation); black bars correspond to $\tau/\tau_R = 0.05$, white bars correspond to $\tau/\tau_R = 2$, and grey bars indicate overlap. The analytic prediction is plotted as a dashed line. (Online version in colour.)

We now derive an expression for the joint pdf of the apparent step lengths and angle changes, R and Θ, denoted $f_{R,\Theta}(r, \theta)$, assuming that the functional form of the angular pdf $f_\Phi(\phi)$ and the value of c_{const} are both known. Knowledge of this pdf permits us to approximate the pdf of RASs and AACs when $\tau/\tau_R \ll 1$. We use stochastic simulations to study the range of values of τ/τ_R for which our approximation holds.

We first define non-dimensional parameters by rescaling the total path length between sampling points, $c_{const}\tau$, as follows:

$$\hat{L} = \frac{L}{c_{const}\tau} \quad \text{and} \quad \hat{R} = \frac{R}{c_{const}\tau}. \tag{5.3}$$

Note that $\hat{R}$ is equivalent to the RAS. The desired joint pdf, denoted $f_{\hat{R},\Theta}(r, \theta)$ following rescaling, is then given by a sum of conditional joint pdfs:

$$f_{\hat{R},\Theta}(r, \theta) = \mathbb{P}(N = 0; \tau, \tau_R)f^{(0)}_{\hat{R},\Theta}(r, \theta) + \mathbb{P}(N = 1; \tau, \tau_R)f^{(1)}_{\hat{R},\Theta}(r, \theta), \tag{5.4}$$

where $f^{(N)}_{\hat{R},\Theta}(r, \theta)$ is the joint pdf of $\hat{R}$ and Θ, conditional on the number of reorientation events, $N \in \{0, 1\}$. Since $N = 0$ corresponds to no change in r and θ, we have

$$f^{(0)}_{\hat{R},\Theta}(r, \theta) = \delta(r - 1)\delta(\theta), \tag{5.5}$$

where δ denotes the Dirac delta function. This result alone

is unremarkable; it is the inclusion of the $N = 1$ case that explains the observed broadening of the distributions of speeds and angle changes with small τ/τ_R. We now use an analytic description of the bijective mapping between the $(\hat{L}, \Phi)$ and $(\hat{R}, \Theta)$ representations of the $N = 1$ trajectory to compute $f^{(1)}_{\hat{R},\Theta}(r, \theta)$:

$$\cos \Phi = \frac{(\hat{R} \sin \Theta)^2 - (1 - \hat{R} \cos \Theta)^2}{(\hat{R} \sin \Theta)^2 + (1 - \hat{R} \cos \Theta)^2} \tag{5.6}$$

and

$$\hat{L} = \frac{1 - \hat{R} \cos \Theta - \hat{R} \sin \Theta}{2(1 - \hat{R} \cos \Theta)}. \tag{5.7}$$

The joint pdf of $\hat{R}$ and Θ, conditional upon $N = 1$, is then

$$f^{(1)}_{\hat{R},\Theta}(r, \theta) = \begin{cases} \dfrac{r}{(1 - r\cos\theta)}f_{\hat{L}}(l(r, \theta))f_\Phi(\phi(r, \theta)), & r \in [0, 1], \\ 0, & \text{otherwise,} \end{cases} \tag{5.8}$$

where the first term is obtained by evaluating the determinant of the Jacobian and we have made use of the independence of L and Φ. For a Poisson process, the distribution of events within an interval, conditional on the number of events, is uniform [26], and hence $f_{\hat{L}}(l) \equiv 1$, $l \in [0, 1]$. Substituting (5.6) and (5.7) into (5.8) yields

$$f^{(1)}_{\hat{R},\Theta}(r, \theta) = \begin{cases} \dfrac{r}{(1 - r\cos\theta)}f_\Phi\left(\arccos\left[\dfrac{r^2 \sin^2\theta - (1 - r\cos\theta)^2}{r^2 \sin^2\theta + (1 - r\cos\theta)^2}\right]\right), & r \in [0, 1], \\ 0, & \text{otherwise.} \end{cases} \tag{5.9}$$

We now integrate the joint pdf (5.4) to obtain the marginal pdfs for $\hat{R}$ and Θ, respectively, denoted by $f_{\hat{R}}(r)$ and $f_\Theta(\theta)$. In order to compare our results with those from a stochastic simulation, we must specify $f_\Phi(\phi)$. We again use the VM distribution with zero mean and angular deviation $\sigma_\delta = 0.4$. The remaining parameters are $\tau_R = 1$ s, $\tau = 0.2$ s and $c_{const} = 1$ m s^{-1}. Using the functional form of the VM distribution (2.1) in equation (5.9), it is not possible to obtain analytic forms for the marginal pdfs. We therefore use Gaussian quadrature to evaluate the pdf numerically.

Figure 10 shows the observed pdfs of RASs and AACs, computed with the stochastic simulation algorithm for $\tau/\tau_R = 0.05$ and $\tau/\tau_R = 2$, overlaid with the result of performing numerical integration of equation (5.9). The stochastic data are filtered to remove transitions in which no reorientations occur (the $N = 0$ case), as these result in a large number of artificial 0 turns, and RASs equal to 1. This is achieved by discarding transitions with an AAC whose magnitude is less than a defined numerical tolerance, which we take to be 1×10^{-6}. Changing this value by two orders of magnitude
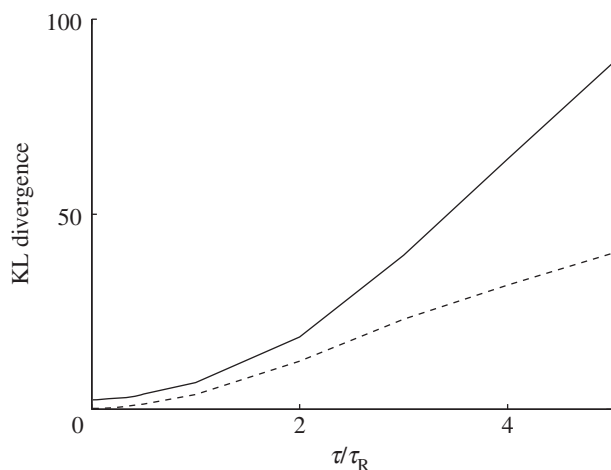
Figure 11. Plot of the variation with the relative sampling interval of KL divergence, defined in equation (5.10), between the results of numerical integration of equation (5.9) and stochastic simulations. Stochastic simulation results are filtered to remove transitions containing no reorientations. Results are shown for the RASs (solid line) and AACs (dashed line).

makes no significant difference to the results (data not shown). The remaining data therefore correspond to one or more reorientation events. The agreement is good when $\tau/\tau_R = 0.05$ as this relative sampling interval is sufficiently small that our earlier assumptions are justified. Conversely, the agreement is poor when $\tau/\tau_R = 2$, with the stochastic simulation generating a broader distribution in both cases. This discrepancy for large relative sampling interval is caused by frequent violation of the underlying assumptions of our analysis.

In order to investigate the range of values of τ/τ_R for which equation (5.9) is valid, we compute the Kullback–Leibler (KL) divergence between the analytic and stochastic results [27]. The KL divergence quantifies the information lost when a discrete pdf, Q, is used to approximate a known discrete pdf, P, and is given by

$$d_{KL} = \sum_x \log\left(\frac{P(x)}{Q(x)}\right) P(x), \qquad (5.10)$$

where x denotes the set of all permissible values of the observed data for which $Q(x) \neq 0$. We let P equal the empirical pdf obtained from a histogram of the stochastic results and evaluate the numerical integrals of equation (5.9) over the same grid spacing to obtain Q. Figure 11 shows the variation of the KL divergence with τ/τ_R for a CRW with $\sigma_\delta = 0.4$. The divergence increases more rapidly with τ/τ_R over the range considered for RASs than for AACs. In both cases, the observed agreement is poor for $\tau/\tau_R \geq 2$ (compare with figure 10). The small residual KL divergence observed when τ/τ_R is very small is because of stochastic fluctuations in P, and decreases as the number of simulation iterations is increased (data not shown).

6. Conclusions

In this paper, we have considered the effect of the sampling frequency on observed quantities of interest arising from an unbiased, constant-speed CRW in two dimensions. We studied two variants of the CRW, denoted the run-only and run-and-stop models. The CRW is a very versatile framework for modelling the motion of organisms. For example, it has

previously been used to incorporate biased motion owing to the influence of gravity and light [19], stationary phases [13] and prokaryotic biochemical signalling pathways [28]. Sampling frequency is an important factor, as it has a strong effect on many of the quantities that researchers are commonly interested in extracting from tracking data [4,12]. For this study, we focused on the two most widely applicable quantities: apparent speeds and angle changes between successive observations. We initially derived results that are valid at long sampling intervals, based on the assumption of a diffusive process. These provided a useful consistency check for the ensuing work. We next analysed intermediate sampling interval regimes using simulated tracks, before using mathematical methods to describe the run-only model when the sampling frequency is high.

In our simulation study, we have extended the work of Codling & Hill [12] by investigating the effect that sampling frequency has on the observed pdf, rather than on common summary statistics. Our work represents a significant advance as summary statistics, while undoubtedly useful information, may hide the complex nature of the sampling effect. For example, figure 4a shows that the distribution of RASs is skewed significantly towards lower speeds, a result not evident by considering the mean and standard deviation alone. Furthermore, several statistical tests based on the empirical cumulative distribution function exist to quantify the compatibility of observed data with a reference distribution [29]; this is not possible with only knowledge of summary statistics. For example, Kareiva & Shigesada [14] develop a method to test whether the observed motion in their tracking data fits a CRW model based on the MSD. The authors assess the compatibility of the data with the CRW by direct comparison of theoretically predicted and empirically observed MSDs, yielding no information relating to the statistical significance of the match. In more recent tracking studies, sufficient data have been available to perform a more in-depth comparison, using the full distribution of speeds and angle changes [30]. We therefore propose that the present study enables a robust test of the statistical compatibility of these observables with the CRW model, taking sampling effects into account. Further work is required to test this conjecture on real tracking data.

The effect of incorporating a stationary reorientation phase into the CRW has not been considered before. This model of motion is applicable to many species that are amenable to study using tracking methods. Examples include bacteria such as *R. sphaeroides* that reorient by stopping the rotation of their flagellum [31], flying insects that pause on landing sites [14] and ruminants that pause at foraging sites [4,17]. We demonstrated that the inclusion of stationary phases in the underlying motion leads to complex variation in the observed standard deviation of RASs, and suggested an explanation based on consideration of the evolution of the RAS pdf. This again highlights the utility of analysing the full pdf of observed quantities.

An open question arising from the simulation study was how to infer the true distribution of angle changes. Since sampling intervals do not in general coincide with reorientation events, we cannot simply 'read off' the result from a histogram of framewise angle changes. This provided the motivation for our analytic study in §5, in which we derived a mathematical description of the sampling process and its effect on the observed data in the limit of rapid sampling.

In previous studies, the analytic forms for the first, second and fourth apparent displacement moments after a specified number of reorientation events have been determined [14,18]. While these results are related to the problem considered here, they are not directly applicable to a study of the effect of sampling frequency, as the independent variable being considered is the number of reorientation events, not the sampling interval. Furthermore, to our knowledge no studies have derived an analytic description of the pdf of AACs. In contrast, our method permits a description of the variation of the full pdfs of both RASs and AACs with the sampling interval for any given pdf of underlying angle changes, provided it is possible to integrate equation (5.9) numerically. Additional investigation showed that our analytic approximation agrees well with simulated data for sampling intervals corresponding to $\tau/\tau_R \lesssim 1$. This is a wider range of values than expected based on consideration of the probability of more than one reorientation event occurring in a given sampling interval. This suggests that our result for the pdf of RASs and AACs conditional on a single reorientation event is similar to the pdfs conditional on a greater number of reorientations, which we discounted for the purposes of this study. As a result, many tracking experiments are sampled sufficiently rapidly that our results are applicable [1,4,9]. Our approach gives important new mathematical insight, and lays the groundwork for describing more complicated motion models.

The CRW model considered in this study is two-dimensional; further work is needed to investigate whether the findings are significantly different in the case of a three-dimensional CRW. This may be a more realistic model for some types of motile behaviour, such as the motion of air- and water-borne organisms. The tracking in these cases may still be two-dimensional owing to technical constraints [19], in which case it is necessary to determine the combined effects of sampling interval and projection of the observed motion onto a plane.

An important next step in this work is to consider the role of noise, which is present in all tracks obtained experimentally. The sources of noise depend on the application, for example, measurement error in acquiring position fixes [4,32] and errors in computing cell centroids in microscope tracking studies [33,34]. Further work is required to determine the complex interplay between experimental noise and the sampling frequency, which is beyond the scope of this study. Several studies have explicitly considered the joint effects of noise and sampling rate on observed tracks, for example [4,34–36]; however, to our knowledge, no studies exist in which these joint effects have been systematically analysed using a CRW movement model. The simulation approach presented in §4 may be extended in a straightforward manner to include noise, for example by modelling measurement error as a Gaussian process [4]; this represents ongoing work.

A further area in which research is required is in more complex CRW models of motility. For example, our assumption that organisms move with a constant speed is not upheld in many species of bacteria [37], mammals [4] and fish [38]. More work is needed to elucidate the effects of sampling rate when organisms move with varying speeds. It remains to be seen whether it is possible to extend the analytic approach developed in the current study to incorporate this effect.

References

1. Berg HC, Brown DA. 1972 Chemotaxis in *Rhodobacter sphaeroides* analysed by three-dimensional tracking. *Nature* **239**, 500–504. (doi:10.1038/239500a0)

2. Bächler E, Hahn S, Schaub M, Arlettaz R, Jenni L, Fox JW, Afanasyev V, Liechti F. 2010 Year-round tracking of small trans-Saharan migrants using light-level geolocators. *PLoS ONE* **5**, e9566. (doi:10.1371/journal.pone.0009566)

3. Codling EA, Plank MJ. 2011 Turn designation, sampling rate and the misidentification of power laws in movement path data using maximum likelihood estimates. *Theor. Ecol.* **4**, 397–406. (doi:10.1007/s12080-010-0086-9)

4. Laube P, Purves RS. 2011 How fast is a cow? Cross-scale analysis of movement data. *Trans. GIS* **15**, 401–418. (doi:10.1111/j.1467-9671.2011.01256.x)

5. Plank MJ, Codling EA. 2009 Sampling rate and misidentification of Lévy and non-Lévy movement paths. *Ecology* **90**, 3546–3553. (doi:10.1890/09-0079.1)

6. Tomkiewicz SM, Fuller MR, Kie JG, Bates KK. 2010 Global positioning system and associated technologies in animal behaviour and ecological research. *Phil. Trans. R. Soc. B* **365**, 2163–2176. (doi:10.1098/rstb.2010.0090)

7. Mitamura H, Arai N, Sakamoto W, Mitsunaga Y, Maruo T, Mukai Y, Nakamura K, Sasaki M, Yoneda Y. 2002 Evidence of homing of black rockfish *Sebastes inermis* using biotelemetry. *Fish Sci.* **68**, 1189–1196. (doi:10.1046/j.1444-2906.2002.00554.x)

8. Carey FG, Scharold JV, Kalmijn A. 1990 Movements of blue sharks (*Prionace glauca*) in depth and course. *Mar. Biol.* **106**, 329–342. (doi:10.1007/BF01344309)

9. Poole P, Sinclair DR, Armitage JP. 1988 Real time computer tracking of free-swimming and tethered rotating cells. *Anal. Biochem.* **175**, 52–58. (doi:10.1016/0003-2697(88)90359-4)

10. Young HC, Reid TG, Randall LA, Lachowsky LE, Foster DJ, Pengelly CJ, Latty T, Reid ML. 2013 Influences of movement behavior on animal distributions at edges of homogeneous patches. *Int. J. Zool.* **2013**, 602845. (doi:10.1155/2013/602845)

11. Codling EA, Plank MJ, Benhamou S. 2008 Random walk models in biology. *J. R. Soc. Interface* **5**, 813–834. (doi:10.1098/rsif.2008.0014)

12. Codling EA, Hill NA. 2005 Sampling rate effects on measurements of correlated and biased random walks. *J. Theor. Biol.* **233**, 573–588. (doi:10.1016/j.jtbi.2004.11.008)

13. Othmer HG, Dunbar SR, Alt W. 1988 Models of dispersal in biological systems. *J. Math. Biol.* **26**, 263–298. (doi:10.1007/BF00277392)

14. Kareiva PM, Shigesada N. 1983 Analyzing insect movement as a correlated random walk. *Oecologia* **56**, 234–238. (doi:10.1007/bf003796950)

15. Bovet P, Benhamou S. 1988 Spatial analysis of animals' movements using a correlated random walk model. *J. Theor. Biol.* **131**, 419–433. (doi:10.1016/s0022-5193(88)80038-9)

16. Cain ML. 1990 Models of clonal growth in *Solidago altissima*. *J. Ecol.* **78**, 27–46. (doi:10.2307/2261034)

17. Fortin D, Morales JM, Boyce MS. 2005 Elk winter foraging at fine scale in Yellowstone National Park. *Oecologia* **145**, 334–342. (doi:10.1007/s00442-005-0122-4)

18. McCulloch CE, Cain ML. 1989 Analyzing discrete movement data as a correlated random walk. *Ecology* **70**, 383–388. (doi:10.2307/1937543)

19. Hill NA. 1997 A biased random walk model for the trajectories of swimming micro-organisms. *J. Theor. Biol.* **186**, 503–526. (doi:10.1006/jtbi.1997.0421)

20. Mardia KV, Jupp PE. 1999 *Directional statistics*, 1st edn. New York, NY: Wiley.

21. Conn AR, Gould NIM, Toint PL. 1987 *Trust-region methods*. Philadelphia, PA: SIAM.

22. Best DJ, Fisher NI. 1979 Efficient simulation of the von Mises distribution. *J. R. Statist. Soc. C* **28**, 152–157. (doi:10.2307/2346732)

23. Benhamou S. 2006 Detecting an orientation component in animal paths when the preferred direction is individual-dependent. *Ecology* **87**, 518–528. (doi:10.1890/05-0495)

24. Alon U, Camarena L, Surette MG, Aguera B, Liu Y, Leibler S, Stock JB. 1998 Response regulator output in bacterial chemotaxis. *EMBO J.* **17**, 4238–4248. (doi:10.1093/emboj/17.15.4238)

25. Duffy KJ, Ford RM. 1997 Turn angle and run time distributions characterize swimming behavior for *Pseudomonas putida*. *J. Bacteriol.* **179**, 1428–1430.

26. Grimmett GR, Stirzaker DR. 2001 *Probability and random processes*, 3rd edn. Oxford, UK: Oxford University Press.

27. Cover TM, Thomas JA. 2006 *Elements of information theory*, 2nd edn. New York, NY: Wiley-Interscience.

28. Xue C, Othmer HG. 2009 Multiscale models of taxis-driven patterning in bacterial populations. *SIAM J. Appl. Math.* **70**, 133–167. (doi:10.1.1.158.9032)

29. Stephens MA. 1974 EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**, 730–737. (doi:10.1080/01621459.1974.10480196)

30. Selmeczi D, Li L, Pedersen LII, Nrrelykke SF, Hagedorn PH, Mosler S, Larsen NB, Cox EC, Flyvbjerg H. 2008 Cell motility as random motion: a review. *Eur. Phys. J. Sp. Top.* **157**, 1–15. (doi:10.1140/epjst/e2008-00626-x)

31. Pilizota T, Brown MT, Leake MC, Branch RW, Berry RM, Armitage JP. 2009 A molecular brake, not a clutch, stops the *Rhodobacter sphaeroides* flagellar motor. *Proc. Natl Acad. Sci. USA* **106**, 11 582–11 587. (doi:10.1073/pnas.0813164106)

32. Frair JL, Fieberg J, Hebblewhite M, Cagnacci F, DeCesare NJ, Pedrotti L. 2010 Resolving issues of imprecise and habitat-biased locations in ecological analyses using GPS telemetry data. *Phil. Trans. R. Soc. B* **365**, 2187–2200. (doi:10.1098/rstb.2010.0084)

33. Bashar MK, Komatsu K, Fujimori T, Kobayashi TJ. 2012 Automatic extraction of nuclei centroids of mouse embryonic cells from fluorescence microscopy images. *PLoS ONE* **7**, e35550. (doi:10.1371/journal.pone.0035550)

34. Li L, Cox EC, Flyvbjerg H. 2011 'Dicty dynamics': *Dictyostelium* motility as persistent random motion. *Phys. Biol.* **8**, 046006. (doi:10.1088/1478-3975/8/4/046006)

35. Jerde CL, Visscher DR. 2005 GPS measurement error influences on movement model parameterization. *Ecol. Appl.* **15**, 806–810. (doi:10.1890/04-0895)

36. Li T, Raizen MG. 2012 Brownian motion at short time scales. *Ann. Phys.* (*Berlin*) **525**, 281–295. (doi:10.1002/andp.201200232)

37. Packer HL, Lawther H, Armitage JP. 1997 The *Rhodobacter sphaeroides* flagellar motor is a variable-speed rotor. *FEBS Lett.* **409**, 37–40. (doi:10.1016/s0014-5793(97)00473-0)

38. Brodersen J, Nilsson PA, Ammitzbøll J, Hansson LA, Skov C, Brönmark C. 2008 Optimal swimming speed in head currents and effects on distance movement of winter-migrating fish. *PLoS ONE* **3**, e2156. (doi:10.1371/journal.pone.0002156)