Single-Trajectory Characterization of Active Swimmers in a Flow.

Gaspard Junot, Eric Clément, Harold Auradou, and Reinaldo García-García

1PMMH, CNRS UMR 7636, ESPCI Paris, Université PSL, 10 rue de Vauquelin, F-75005 Paris, France
2Université Paris-Saclay, CNRS, FAST, 91405, Orsay, France

We develop a maximum likelihood method to infer relevant physical properties of elongated active particles. Using individual trajectories of advected swimmers as input, we are able to accurately determine their rotational diffusion coefficients $D_R$ and Betherton parameters $\beta$, also providing reliable estimators for the uncertainties of these observables. We test our theoretical construction using numerically generated active trajectories upon no-flow, Couette flow, and Poiseuille flow conditions, with excellent results. Being designed to rely on single-particle data, our method eases applications in experimental conditions where swimmers exhibit a strong morphological diversity. We briefly discuss some of such ongoing experimental applications, specifically, in the characterization of swimming E. Coli in a flow.

I. INTRODUCTION

To monitor passive probes, living cells and microorganisms moving in their environment, modern instrumentation is now offering possibilities for long-time tracking and trajectory reconstruction, with excellent spatial and temporal resolution [1]. The collected data sets contain in principle relevant informations on the internal and external processes, of either mechanical or biological origin. However, inferring them is still a challenging task which has motivated the recent development of adapted analytical tools [2–7].

Tracking colloids close to thermal equilibrium media was proven quite fruitful as it allowed the development of a quantitative micro-rheology instrumentation [8]. This technique was further extended to investigate complex fluids [9, 10], and dynamical processes inside living cells [1, 11, 12]. In recent years, much effort has been dedicated to the study of passive colloidal particles as they diffuse in active suspensions composed of self-propelling entities [13–15]. Several studies focus on the contribution of the active environment to the diffusion of tracers [16–18], while others aim at generalizing the notion of thermal bath and to characterize the thermodynamic fluctuations associated to the passive particle in complex environments [19–22]. However, there has been much less focus on understanding how an active particle gets itself affected by the properties of its the environment.

In this work, we develop a maximum likelihood (ML) approach to study elongated active swimmers using raw data of swimming tracks. The general aim is to extract the physical parameters that characterize the particle from stochastic data. In general, such data encompass the active transport contribution resulting from the interaction between the flow and the swimmer, and the stochastic component associated with rotational noise. Although our framework is general, we focus here on three archetypical examples: a free swimmer, a swimmer in a shear flow and, finally, a swimmer in a Poiseuille flow. We take into account some limitations encountered in experiments, such as the finite duration of the tracks and their discrete sampling. The outcome of the method is tested against numerical simulations, which show an excellent agreement between prescribed parameter values and our estimators.

Our tool was originally conceived to study E. Coli mutants (smooth-runners) swimming in a parabolic flow, but it is easily generalizable to different experimental scenarios. Smooth swimmers were recently considered in Ref. [23]. It was shown that a deterministic model, consisting of an advected swimming ellipsoid, fairly reproduces bacterial trajectories at short times. However, at longer times, a stochastic component comes into play as a multiplicative noise leading to the stochastic exploration of the phase space. Using our method, we are able to extract the effective rotational diffusivity from the swimming trajectories, which can be compared quantitatively with the Brownian diffusion of an ellipsoid in a viscous fluid. Along these lines, we report on a preliminary analysis of experimental trajectories which further supports the accuracy of our method.

The remainder of the paper is organized as follows. In Sec. II we introduce the theoretical model and briefly discuss the philosophy of our ML method. In Sec. III, we derive our ML estimators and their uncertainties to leading order in the number of sampling points. In Sec. IV we use numerical simulations to validate our method in three flow configurations, namely, no-flow, Couette flow and Poiseuille flow. A preliminary application of our method in the analysis of experimental tracking data is discussed in Sec. V, and general conclusions are discussed Sec. VI. Finally, all technical details regarding our calculations and experiments are left for the Appendix.

II. PRELIMINARIES

A. Stochastic dynamics of the active Betherton-Jeffery model

Smooth swimmers are affected by environmental fluctuations. This represents a strong motivation to explicitly analyze noise effects in the rotational dynamics of the orientation vector $p$ of swimming bacteria. We focus on
the model studied in Ref. [23] in presence of noise, which we refer to as the stochastic active Betherton-Jefferey (SABJ) model from now on:
\[ \dot{r} = v_0 \mathbf{p} + \mathbf{v}^F(r), \quad (1) \]
\[ \dot{p} = (\mathbb{1} - p \otimes p) [\beta \mathbb{E}(r) + \Omega(r)] p - 2D_R p + \sqrt{2D_R} p \wedge \xi_R, \quad (2) \]

Above, \( v_0 \) is the self-propulsion velocity of the particle, assumed constant, \( \mathbf{v}^F \) is the local flow velocity, the symbol \( \mathbb{1} \) denotes the identity matrix, and \( \otimes \) is used to denote tensor products. The number \( \beta = (r^2 - 1)/(r^2 + 1) \) is the Betherton parameter, which represents a measure of the geometrical asymmetry of the swimmer (\( r \) is the aspect-ratio of the ellipsoidal particle). The components of the tensors \( \mathbb{E} \) and \( \Omega \) are given as follows, \( E_{ij}(r) = (\partial_{r_i} v^F_{j} + \partial_{r_j} v^F_{i})/2 \) and \( \Omega_{ij}(r) = (\partial_{r_i} v^F_{j} - \partial_{r_j} v^F_{i})/2 \).

In Eq. (2), rotational diffusion is encoded in the diffusion coefficient \( D_R \) and the Gaussian white noise \( \xi_R \) which has zero mean and variance \( \langle \xi_R(t) \otimes \xi_R(t') \rangle = 2D_R \delta(t - t') \). In this work, Eq. (2) is interpreted in the Ito sense. The Ito term \(-2D_R \mathbf{p}\) is thus needed to guarantee the conservation of the norm of \( \mathbf{p} \). Note that, although one could also introduce a translational diffusion term in Eq. (1), such contribution can be assumed negligible in typical experimental situations. Consider, for instance, an \( E. \text{Coli} \) bacterium which is few microns length, swimming at a typical speed of 25 \( \mu \text{m} \text{s}^{-1} \). The distance over which it diffuses in 1 s is \( \approx 0.1 \mu \text{m} \) to be compared to the 25 \( \mu \text{m} \) travelled due to its activity.

An important role will be played, in what follows, by the Markov kernel of the process (1)-(2), which we denote by \( K(\mathbf{r}, \mathbf{p}, t|\mathbf{r}' \mathbf{p}', t') \). We focus here on stationary flow profiles in which case the propagator only depends on \( \Delta t = t - t' \) due to time-translation invariance. We thus simplify the notations by writing \( K(\mathbf{r}, \mathbf{p}, t|\mathbf{r}', \mathbf{p}', t') = K(\mathbf{r}, \mathbf{p}, \Delta t|\mathbf{r}', \mathbf{p}', 0) \equiv K(\mathbf{r}, \mathbf{p}, \Delta t|\mathbf{r}', \mathbf{p}') \). The propagator satisfies the Fokker-Planck equation \( \partial_t K = \hat{L}_{FP} K \) with initial condition \( K(\mathbf{r}, \mathbf{p}, 0|\mathbf{r}', \mathbf{p}') = \delta(\mathbf{r} - \mathbf{r}') \delta(\mathbf{p} - \mathbf{p}') \). The Fokker-Planck operator associated to these dynamics reads
\[ \hat{L}_{FP}(\mathbf{r}, \mathbf{p}) = D_R \sum_{i,j} \partial^2_{p_i p_j} M_{ij}(\mathbf{p}) - \sum_i \partial_{p_i} \cdot h_i(\mathbf{r}, \mathbf{p}) + \sum_i \partial_{x_i} \cdot v_i(\mathbf{r}, \mathbf{p}). \quad (3) \]

To write (3), we condensed the translational part of the motion into \( \mathbf{v}(\mathbf{r}, \mathbf{p}) = v_0 \mathbf{p} + \mathbf{v}^F(\mathbf{r}) \), and the drift term in Eq. (2) into \( h(\mathbf{r}, \mathbf{p}) = (\mathbb{1} - \mathbf{p} \otimes \mathbf{p}) [\beta \mathbb{E}(r) + \Omega(r)] \mathbf{p} - 2D_R \mathbf{p} \). Additionally, each element of the matrix \( \mathbb{H}(\mathbf{p}) \) is a quadratic form in \( \mathbf{p} \), \( M_{ij}(\mathbf{p}) = \sum_{l,q} T_{ijql} p_l p_q \), where
\[ T_{ijql} = \sum_k \epsilon_{ilk} \epsilon_{jql} = \delta_{ij} \delta_{lq} - \delta_{iq} \delta_{jl}, \quad (4) \]
and \( \epsilon_{ijk} \) denotes the Levi-Civita symbol.

Solving exactly the Fokker-Planck equation for the propagator of the SABJ model in presence of generic flows is a difficult task. As this Markov kernel plays an important role in our discussion, we develop a reasonably good approximation for it. Our approximation is a key element of the inference method we develop in what follows.

### B. Estimating parameters from single trajectories

Following the conclusions of Ref. [23], one may argue that the SABJ model provides a faithful representation of the dynamics of smooth swimmers far from the walls of the container. Here we built on this hypothesis. Our main goal is to be able to rely on experimental data to address, for instance, the variability of physical traits in bacteria. By studying individual trajectories of many different swimmers, one could access the statistics of their diffusion coefficients. It is then crucial to be able to accurately estimate the parameters of the SABJ model from individual tracks.

To achieve our goal, we develop an ML method. We start by constructing a good approximant for the log-likelihood characterizing a single-bacterium trajectory; then the relevant physical parameters are determined by maximizing it. To be more specific, consider an experimental track of a swimming bacterium. Continuous trajectories are sampled at regularly spaced time instants \( \{ t_\alpha \} \), such that \( \Delta t_\alpha = t_{\alpha+1} - t_\alpha \equiv f^{-1} \) for each \( \alpha = 0, 1, \ldots, N - 1 \), where \( f \) is the sampling frequency. The data has the form of a discrete set of \( N \) position and orientation values, \( \{ \mathbf{r} \} \). Assuming that the dynamics of the swimmer are compatible with Eqs. (1)-(2), the probability of measuring that particular track can be expressed as follows:
\[ P(\{ \mathbf{r} \}) = \prod_{\alpha=0}^{N-1} K(\mathbf{r}(t_{\alpha+1}), f^{-1}|\mathbf{r}(t_\alpha)). \quad (5) \]

Although not made explicit, the probability of the track is parametrized by the relevant physical quantities of the model, i.e., \( v_0, D_R, \) and \( \beta \). One can introduce the log-likelihood as \( S = \ln P \), or more explicitly
\[ S = \sum_{\alpha=0}^{N-1} \ln K(\mathbf{r}(t_{\alpha+1}), f^{-1}|\mathbf{r}(t_\alpha)). \quad (6) \]

anticipating why the Markov kernel of the SABJ model is important in this work. By maximizing the log-likelihood with respect to the parameters, we obtain the best estimates compatible with the hypothesis that the data are generated by Eqs. (1)-(2).

We remark that our procedure can be formally justified by the fact that the SABJ model is Markovian, which means that the propagator encodes all the relevant physical information about the process. Markovianity, in conjunction with time-translation invariance, allows us then to interpret one trajectory as an ensamble of independent ‘events’, each one consisting on the generation
of new values for the increments of the process, with each outcome being fully characterized by the same probabilistic model $K$.

An important remark that needs to be made, is that our method can be used to infer parameters from tracking data using generic models and not only the one we consider here. As far as time-translation invariance and Markovianity hold, it all comes down to compute the corresponding propagator and following the general steps sketched above. In the next section, we develop in detail the ML method for estimating the parameters of the SABJ model from individual experimental tracks. We will, in particular, pay special attention to some technical subtleties that need to be addressed with care.

III. MAXIMUM LIKELIHOOD METHOD

A. Approximate log-likelihood and maximum likelihood estimates

As mentioned before, solving the Fokker-Planck equation for the propagator in presence of a generic flow seems hopeless. This implies that, in practice, the likelihood cannot be constructed explicitly in an exact manner. Nevertheless, a fairly good approximant can be derived.

Consider the (ideal) limit at which the trajectory of the swimmer is sampled continuously, corresponding to $f \to \infty$. In that case the track probability (5) becomes the continuous path probability associated to the process (1)-(2). It is then natural, if the sampling frequency is large enough but finite, to approximate the time integrals involved in the expression of the continuous path probability by finite sums, exactly like approximating Riemann integrals by Darboux sums with a sufficiently small time step. Using this idea, one can derive an approximate expression for the probability of a discrete sample of points belonging to a trajectory generated by the model (1)-(2) (see Appendix A). We now use this result (c.f. Eq. (A14)) to write down the approximation used in this work for the log-likelihood.

Let us introduce the notations $\Delta r_\alpha = r_{\alpha+1} - r_\alpha$, $\Delta p_\alpha = p_{\alpha+1} - p_\alpha$, $a_\alpha = f \Delta p_\alpha - (1 - p_\alpha \otimes p_\alpha) \partial (r_\alpha) p_\alpha$, and $b_\alpha = (1 - p_\alpha \otimes p_\alpha) \partial (r_\alpha) p_\alpha$, so that we have $f \Delta p_\alpha - h_\alpha = a_\alpha - b_\alpha a_\alpha + 2D_R p_\alpha$. We also define

$$S_\alpha = p_{x,\alpha}^2 + p_{z,\alpha}^2,$$

$$A_\alpha = a_{z,\alpha} p_{x,\alpha} - a_{x,\alpha} p_{z,\alpha},$$

$$B_\alpha = b_{z,\alpha} p_{x,\alpha} - b_{x,\alpha} p_{z,\alpha},$$

$$C_\alpha = a_{y,\alpha} S_\alpha - (a_{x,\alpha} p_{x,\alpha} + a_{z,\alpha} p_{z,\alpha}) p_{y,\alpha},$$

$$E_\alpha = b_{y,\alpha} S_\alpha - (b_{x,\alpha} p_{x,\alpha} + b_{z,\alpha} p_{z,\alpha}) p_{y,\alpha},$$

$$A_R = f^{-2} \sum_\alpha S_\alpha^{-1} (A_\alpha^2 + C_\alpha^2),$$

$$B_R = f^{-2} \sum_\alpha S_\alpha^{-1} (A_\alpha B_\alpha + C_\alpha E_\alpha),$$

$$C_R = f^{-2} \sum_\alpha S_\alpha^{-1} (B_\alpha^2 + E_\alpha^2).$$

(7)

The log-likelihood can be derived by taking the logarithm of Eq. (A14) after having into account that $|p_\alpha| = 1$ for all $\alpha$ in any experimental or numerical track, and that Eq. (1) is used to fix the evolution of the vector $p$. To take the logarithm, we need first to adimensionalize the probability. We then define $\hat{p} = \pi^{-N} \mathcal{P}_C$, with $\mathcal{P}_C$ given by (A14). It is also convenient to introduce the dimensionless inverse diffusion coefficient $\hat{N}_R = D_R^{-1} f$. With all this, we can finally write $S = \ln \hat{p}$, which gives

$$S(\hat{N}_R, \beta) = N \ln \hat{N}_R - \frac{\hat{N}_R}{4} (A_R - 2B_R \beta + C_R \beta^2).$$

(8)

Maximizing $S$ as given by Eq. (8) is a simple exercise. Assuming that $\hat{N}_R$ is different from 0, we obtain the following estimates

$$\beta = \frac{B_R}{C_R},$$

(9)

$$\hat{N}_R = \frac{4NC_R}{A_RC_R - B_R^2}.$$  

(10)

For consistency, all these results must be positive numbers. We discuss below on the accuracy of such estimates.

B. Note on convergence and uncertainties

From a methodological point of view, it is crucial to be able to quantify the quality of the estimates (9) and (10). A first important question is whether our approximation for the log-likelihood is compatible with experimental data or not. In other words, we need to provide a quantitative meaning to the statement “$f$ is sufficiently large”. A good way to answer this question is to have a closer look to the consistency condition derived in Appendix A, $p_\alpha \cdot \Delta p_\alpha + 2D_R f^{-1} |p_\alpha|^2 = 0$. As discussed in the Appendix, this condition is linked to the conservation of the norm of $p$ but, as we immediately show, this link is broken if $f$ is not large enough.

To understand this, let us consider the scalar product $p_\alpha \cdot p_{\alpha+1}$ for vectors with constant norm, $|p_\alpha| = |p_{\alpha+1}| = 1$ for all $\alpha$. From one side, we have $p_\alpha \cdot p_{\alpha+1} = \cos(\phi_\alpha)$,
where $\phi_\alpha$ is the angle between $\mathbf{p}_\alpha$ and $\mathbf{p}_{\alpha+1}$. On the other hand, we can write $\mathbf{p}_\alpha \cdot \mathbf{p}_{\alpha+1} = \mathbf{p}_\alpha \cdot (\mathbf{p}_\alpha + \Delta \mathbf{p}_\alpha) \equiv 1 + \mathbf{p}_\alpha \cdot \Delta \mathbf{p}_\alpha$. Using then the condition derived in the Appendix, we finally get $\mathbf{p}_\alpha \cdot \mathbf{p}_{\alpha+1} = 1 - 2D_R f^{-1}$, which implies $\cos(\phi_\alpha) = 1 - 2D_R f^{-1}$. It is now clear that our approximation breaks down when $f \leq D_R$, since in that case, one would have $|\cos(\phi_\alpha)| > 1$. In other words, our approximation is not reliable when the rotational diffusion coefficient to be estimated is larger than the sampling frequency, which means that if the estimated $D_R$ ends up being larger than the sampling frequency $f$, such result should not be trusted.

We now focus on the efficiency of the method. More precisely, we estimate the speed at which the computed parameters converge toward their exact values as a function of the length of the sample, $N$. Let us denote by $\theta = (\beta, \hat{N}_R)$ the estimated two-dimensional parameter vector, and by $\Delta \theta$ the estimation error. We also write $\theta^*$ to denote the vector of the true values of the parameters. If the assumption that the data is generated by the model (1)-(2) is statistically correct, then one can assert that, up to order $N^{-1/2}$, the estimated parameters are unbiased. More precisely, the error vector $\sqrt{N}\Delta \theta$ converges in distribution to a normal distribution of zero mean and variance matrix $\bar{F}^{-1}(\theta^*)$, where $\bar{F}(\theta^*)$ is the Fisher information matrix [24], which can be approximately determined from the log-likelihood (for large $N$) as $F_i(\theta^*) = -N^{-1} \partial_\theta^2 \bar{G} |_{\theta = \theta^*}$. Beyond order $N^{-1/2}$, the estimates (9) and (10) are known to exhibit a bias of order $O(1/N)$, but we can neglect such contributions if $N$ is large enough. From this calculation, we estimate the following error bars (we use the notation $e(\theta_i) = \sqrt{\langle (\theta_i^2) \rangle}$):

$$
e(\beta) \sim \frac{2\sqrt{N}}{\sqrt{\hat{N}_R^* (2NC_R - \hat{N}_R^* (B_R - \beta^* C_R)^2)]}, \quad (11)$$

$$
e(\hat{N}_R) \sim \frac{\sqrt{2C_R \hat{N}_R^*}}{\sqrt{2NC_R - \hat{N}_R^* (B_R - \beta^* C_R)^2}}} \quad (12)$$

Interestingly, there are also cross-correlations between the errors, meaning that $\beta$ and $\hat{N}_R$ cannot be determined independently with arbitrarily high accuracy. Explicitly, we have

$$\langle \Delta \beta \Delta \hat{N}_R \rangle \sim \frac{2\hat{N}_R^* (B_R - \beta^* C_R)}{2NC_R - \hat{N}_R^* (B_R - \beta^* C_R)^2]. \quad (13)$$

As a consistency check, note that the quantities $A_R$, $B_R$ and $C_R$ are extensive for $N \gg 1$ ($Z_R \sim \infty$, with $Z = A, B, C$). We then have $e(\theta) \sim N^{-1/2}$, as it should be.

IV. NUMERICAL VALIDATION OF THE METHOD

In the previous section, we derived ML estimators for the parameters of the SABJ model. To test these ex-

pressions and their accuracy, we use the ML method on simulated trajectories, as explained in the following.

Bacterial trajectories are simulated using Eqs. (1)-(2), with an Euler-Maruyama scheme and with a given set of parameters $\beta^*$ and $\hat{N}_R$. Then, we apply the ML method (Eq. (9) and Eq. (10)) on these trajectories to estimate the parameters $\beta$ and $\hat{N}_R$. Finally, we compare our estimators with the input parameters of the simulations.

The ML method is general and can be applied to any stationary flow profile. We choose here three different cases. First, we consider a free swimmer, in which case only the diffusion coefficient can be estimated. Note that in that scenario the parameter $\beta$ plays no role and, correspondingly, it does not appear in the dynamic equations. We then test the method further using two additional flow profiles: a simple shear flow and a Poiseuille flow. In all cases, the convergence of the estimators as function of the length of the simulated trajectory is analyzed.

A. Free swimmer

Let us consider a bacterium swimming in an unbounded environment without flow. Its orientation follows a diffusion process on the unit sphere:

$$\dot{\mathbf{p}} = -2\mathbf{p} + \sqrt{2}\mathbf{p} \wedge \mathbf{\xi}_R.\quad (14)$$

Here, we have adimensionalized time using the rotational diffusion coefficient $D_R$. We simulate trajectories with different duration $T$. The time interval between two sampling events is fixed as $\Delta t = 10^{-6}$. To test the convergence of the method, we compute the diffusion coefficient $D_{R,M}$, using eq (10), for trajectories of different duration. The longer the trajectory, the better the estimation is. In Fig. 1, we illustrate the convergence of the estimated value $D_{R,M}$ towards the prescribed value $D_R = 1$ as $T$ increases. For instance, we find a value of $D_{R,M}$ very close to $1$, i.e $D_{R,M} = 0.9994$, at $T = 1$. 

![Fig. 1. Convergence of the estimated rotational diffusion coefficient $D_{R,M}$ in the free swimmer case. Inset: zoom close to the region $D_{R,M} = 1$.](image-url)
B. Shear flow

In this case we simulate Eqs. (1) and (2) with a flow profile of the form \( \mathbf{v}^F(\mathbf{r}) = \frac{\gamma}{2} \mathbf{e}_x \), where \( \gamma \) is the shear rate. As in the free swimmer case, we work with dimensionless variables. Thus, in addition to \( \beta \), the relevant dimensionless parameter to be estimated in this case is the Péclet number, \( Pe = \frac{\gamma T}{D} \).

We present the results of the method on simple shear flow simulations. For a given trajectory length \( T \), the estimated value of the parameters (for both \( \beta \) Fig. 2(a) and \( Pe \) Fig. 2(b)) are scattered around the input parameter values. This scattering is linked to the randomness encoded in the noise and the initial orientations. The average value of the estimations over all trajectories of the same length (dark blue points) is very close to the input value of the simulations. On average, even for short trajectories, the ML method gives a good estimation of the parameters. For instance, for a trajectory of duration \( T = 10 \), the averaged estimated value differs only of 1% from the input values. As \( T \) increases, the scattering of the estimated values around the input values decreases and our estimators are accurate even for individual trajectories. Moreover, the standard deviation of the estimated parameters is bounded by the uncertainties computed using the expressions Eq. (11) and Eq. (12) and goes as \( N^{-1/2} \).

C. Poiseuille flow

We now test the method on a Poiseuille flow profile. Here we considered a flow confined between two infinite plates separated by a vertical distance \( h \), so that the velocity profile is given by \( \mathbf{v}^F(\mathbf{r}) = \frac{4U_m}{h} z \left( 1 - \frac{z}{h} \right) \mathbf{e}_x \). \( U_m \) denotes the maximal flow velocity at the middle of the Poiseuille flow (at \( z = h/2 \)). We work with dimensionless variables, so that lengths and time are adimensionalized using \( h \) and the maximum shear rate \( \dot{\gamma}_M = \frac{4U_m}{h} \), respectively. In addition to \( \beta \), the relevant dimensionless parameter is again the Péclet number, \( Pe = \frac{\gamma T}{D} \). We simulate trajectories for a set of parameters: \( (\beta^*, Pe^*) = (0.9, 100) \) and of different duration \( T \).

In Fig. 3, we present the corresponding results. In the Poiseuille flow case, a difficulty arises. We now have to consider the borders of the channel (in \( z = 0 \) and \( z = 1 \)), so when a trajectory touches the borders, the simulation ends and we can not have, as it was the case in the simple shear flow, trajectories of arbitrary length nor a prediction of the trajectory duration. For the parameters used in the simulations, it is then difficult to have trajectories longer than \( T = 50 \). Nevertheless, the results are qualitatively similar to those in the simple shear flow case.

For a given \( T \), the estimated values are scattered around the input value. As \( T \) increases the scattering decreases. The estimation of the Pe number remains very good, both on average (less than 2% of error) and for individual trajectories (less than 10% of error), even for trajectories with durations as short as \( T = 1 \). The estimation of \( \beta \) is not as good as in the shear flow, but we still have an average estimated value at 10% of the input one. The standard deviation of the estimated values are still bounded by the expressions Eq. (11) and Eq. (12) and still goes as \( N^{-1/2} \). We expect these estimations to improve further by computing errors up to \( O(N^{-1}) \).
these bacteria aim at swimming in a straight line, their orientation slowly diffuses due to thermal agitation. Further details on the experimental setup are given in the Appendix.

We first consider the free swimmer case. The absence of flow makes the determination of the rotational diffusion coefficient easier and standard methods can be applied. For each 3D bacterial trajectory, we computed the correlation function of the orientation vector \( \mathbf{p} \) as a function of time, which decays exponentially with a characteristic time scale \( \tau = \frac{1}{2D_R} \) with \( D_R \) the rotational diffusion coefficient. This provides a straightforward way of determining the rotational diffusion coefficient for each bacterium.

To be able to use the ML method, the time interval between two samplings, \( \Delta t \) has to fulfill some conditions, as discussed in Sec. III B. First, it has to be small enough so that our approximations are reliable. However, it should not be too small either. Indeed, in experimental situations, the measurements carry some implicit noise. If \( \Delta t \) is too small, the recorded change in the bacterium orientation between any instant \( t \) and \( t + \Delta t \) will be dominated by this noise and not by real physical contributions. Experimental trajectories, which are sampled at a frequency \( f_{ac} \), must then be resampled accordingly, so that in any time window the experimental noise is reduced.

The rotational diffusion coefficient, \( D_R \), characterizes the Brownian motion of the orientation of an ellipsoidal object. It must thus be compared to the estimate [25, 26]

\[
D_B = \frac{3k_B T l n(2l/a)}{\pi \eta l^3}
\]  

![FIG. 3. Convergence of the estimated (a) \( \beta \) and (b) Pécel parameters. Ratio of the estimated value (\( \beta^M \) or \( Pe^M \)) and the input value (\( \beta^* \) or \( Pe^* \)) as function of the trajectory duration \( T \). For each duration \( T \), 20 trajectories were simulated. Each cyan blue point corresponds to the estimation of one trajectory, the dark blue points are the average of these estimations over the 20 trajectories with corresponding standard deviation. The red curves represent the uncertainties computed using the expressions Eq. (11) (resp. Eq. (12)), i.e. are the curves of equation: \( y = 1 \pm e(\beta) \) (resp. \( y = 1 \pm e(Pe) \)).](image)

V. EXPERIMENTAL DETERMINATION OF THE PARAMETERS

In this section, we give an example of the applicability of the ML method on experimental bacterial trajectories. First, we check that the ML well reproduces results that could be obtained with standard methods. To do so, we recorded 3D trajectories of bacteria confined between two glass slides, separated by a distance \( h = 260 \mu m \). The acquisition frequency is fixed to the value \( f_{ac} = 100 \text{ Hz} \). For this experiment, we chose a mutant strain of *E. Coli* (CR20) for which tumbling events have been genetically inhibited (smooth swimmers). Therefore, while obtained by Perrin, where \( l \) is the length of the ellipsoid and \( a \) its width, while \( \eta \) denotes the viscosity of the surrounding medium at a temperature \( T \). For an elongated object of length \( l = 10 \mu m \) and width \( a = 1 \mu m \) (typical bacterium dimensions), it gives \( D_B \approx 0.012 s^{-1} \). One should then have \( f \gg D_B \), accordingly we chose \( 1/f = \Delta t = 1/3 s \). From a trajectory \( \Gamma(1/f_{ac}) \), we consider a resampling \( \Gamma(1/f) \) by taking one data point each \( \Delta t \). By moving the first taken point, we obtain from one trajectory \( \Gamma(i/f_{ac}) \), a set of trajectories \( \Gamma_i(i/f) \), \( i = 1, ..., N \), \( N = f_{ac}/f \). To estimate the rotational diffusion coefficient of one experimental trajectory, we apply the ML method on each resampled trajectory \( \Gamma_i \) of the set. We obtain \( N \) values of the estimated \( D_R \) that we average to get \( D^M \), the ML estimation of the rotational diffusion coefficient of one bacterial trajectory. We then applied this method on the pool of trajectories on which we computed \( D_R \) with the decorrelation method. Fig. 4(a) displays a scatter plots where \( D_R \), the rotational diffusion coefficient computed using the decorrelation, is in the vertical axis, while the horizontal axis corresponds to \( D^M \), the one computed using the ML method. One can immediately see that the ML estimation works remarkably well.

We then proceed to test the ML method in presence
of an external flow. To do so, we inject a suspension of smooth swimmers inside a microchannel of rectangular cross section (height $h = 100 \, \mu m$ and width $w = 600 \, \mu m$). Far from the lateral walls, one thus has a Poiseuille flow. Dozens of 3D bacterial trajectories are recorded at different flow rates $Q$, ranging from 1 to 4.3 $nL/s$. Here we focus on bulk trajectories that are at least 10 $\mu m$ away from the top and the bottom walls and that last at least $10\Delta t$ ($\Delta t = 1/3 \, s$). We then apply the ML method to those trajectories to extract the Péclet number. Fig. 4(b) displays the results of the estimation of $Pe$ on experimental bacterial trajectories. One can see that $Pe_M$ increases with $\dot{\gamma}_M$, as it is expected for a constant rotational diffusion coefficient. By a linear fit on the average values of $Pe$, we get the rotational diffusion coefficient of the bacteria which is the inverse of the slope. It then gives $D_M = 0.032 \pm 0.016 \, s^{-1}$. This value is consistent with the value $< D_r^M > = 0.029 \, s^{-1}$ measured in absence of flow, as well as with values reported in presence of flow [23].

VI. DISCUSSION

In this article, we have developed a ML method suitable to infer parameters from individual stochastic trajectories. We have been able to estimate the parameters of the SABJ model accurately using both, numerical simulations and experiments. The accuracy of our estimators based on numerical data is excellent. The preliminary results we obtained from the analysis of experimental tracks are encouraging. One of the outcomes of our numerical study is that uncertainties remain of order $O(N^{1/2})$ even for relatively short trajectories. These results can be further improved by considering errors up to $O(N^{-1})$, in which case our estimators acquire a bias.

Our estimators for the rotational diffusion coefficient are more scattered in experiments than in simulations. This can be understood by noting that the shape of a bacterium affects the value of $D_R$, which depends sensitively on the effective value of the ellipsoid long axis $l$. For an effective ellipsoid of length $l = 5 \div 10 \, \mu m$ and width $a = 1 \, \mu m$, the estimation of $D_R$ using Eq. (15), yields $D_R = 1.25 \div 5.10^{-3} \, s^{-1}$. We thus have an intrinsic variation of $D_R$ in a bacterial population.

Our ML method is general and can be applied to any Markovian and time-translationally invariant stochastic process. In the context of swimming bacteria, one can study variations of parameters due to external conditions and physical effects that are not considered in the SABJ model. For instance, shear may shorten the flagella bundle leading to higher effective $D_R$ (i.e lower $Pe$). Modifying the SABJ model as to account for rheotactic effects [28, 29], and adapting our ML method appropriately, one could study rheotactic drift [27] in detail, analyze possible changes in bundle formation due to shear, and probe the flexibility of the hook at the base of the flagellum and its interplay with external flow. These and other physical questions will be addressed in the future.

ACKNOWLEDGMENTS

This work was supported by the ANR grant “BacFlow” ANR-15-CE30-0013. This work received the support of Institut Pierre-Gilles de Gennes (Équipement d’Excellence, Investissements d’Avenir program ANR-10-EQPX-34). RGG acknowledges support from ANR grant ANR-17-CE08-0047-02.
Appendix A: Discretized path-integral representation of the SABJ model

Our starting point is the Martin-Siggia-Rose-de Dominics-Janssen (MSRDJ) path integral representation of the process (1)-(2) (for details on the method see Refs. [30–32]):

\[ \mathcal{P}[\Gamma(t)] = \mathcal{J}[\Gamma(t)] \int D[\tilde{\Gamma}(t)] e^{-\mathcal{A}[\tilde{\Gamma}(t),\hat{\Gamma}(t)]}, \]  

(A1)

\[ \dot{\mathcal{A}}[\Gamma, \tilde{\Gamma}] = \frac{1}{f} \sum_{\alpha} \left\{ \left[i f \Delta r_{\alpha} - v_{\alpha}\right]^T \dot{r}_{\alpha} + i \left[f \Delta p_{\alpha} - h_{\alpha}\right]^T \dot{p}_{\alpha} + D_R \dot{p}_{\alpha}^T \tilde{v}(p_{\alpha}) \dot{p}_{\alpha} \right\}, \]  

(A3)

where \( \Gamma_{\alpha} (\hat{\Gamma}_{\alpha}) \) is a shorthand notation for \( \Gamma(t_{\alpha}) (\hat{\Gamma}(t_{\alpha})) \), while \( \Delta \Gamma_{\alpha} \equiv \Gamma_{\alpha+1} - \Gamma_{\alpha} \). Additionally, we have introduced the notations \( v_{\alpha} \equiv v(\Gamma_{\alpha}) \) and \( h_{\alpha} \equiv h(\Gamma_{\alpha}) \). With all this, we write

\[ \mathcal{P}([r, p] = \int \prod_{\alpha} d^3 r_{\alpha} d^3 p_{\alpha} \exp \left( -\frac{1}{f} \sum_{\alpha} \left\{ \left[i f \Delta r_{\alpha} - v_{\alpha}\right]^T \dot{r}_{\alpha} + i \left[f \Delta p_{\alpha} - h_{\alpha}\right]^T \dot{p}_{\alpha} + D_R \dot{p}_{\alpha}^T \tilde{v}(p_{\alpha}) \dot{p}_{\alpha} \right\} \right). \]  

(A4)

The integrals over \( \{\dot{r}_{\alpha}\} \) are immediate and can be readily performed to yield

\[ \mathcal{P}([r, p] = \prod_{\alpha} \delta(\Delta r_{\alpha} - f^{-1} v_{\alpha}) \int \prod_{\alpha} d^3 p_{\alpha} \exp \left( -\frac{1}{f} \sum_{\alpha} \left\{ \left[i f \Delta p_{\alpha} - h_{\alpha}\right]^T \dot{p}_{\alpha} + D_R \dot{p}_{\alpha}^T \tilde{v}(p_{\alpha}) \dot{p}_{\alpha} \right\} \right). \]  

(A5)

The remaining integrals over \( \{p_{\alpha}\} \) are less trivial because, as can be readily verified, the matrix \( \tilde{v} \) is singular, i.e., \( \text{det}(\tilde{v}) = 0 \). Such singularity is linked to the fact that the continuous dynamics Eq. (2) imposes a hard constraint on \( p \), i.e., \( |p|^2 = 1 \). This means that at any time instant only two numbers are needed to fully describe the state of the three-dimensional vector \( p \). For instance, by knowing the projection of \( p \) on the \( z \) axis, \( p_z \), and the polar angle \( \theta \) of the projection of \( p \) in the \( x - y \) plane, one can write \( p = (\sqrt{1 - p_z^2} \cos \theta, \sqrt{1 - p_z^2} \sin \theta, p_z) \).

From this discussion, one can infer that there is a relevant two-dimensional sub-space allowing to fully describe \( p \). One can indeed isolate the singularity, so that a Gaussian integration can still be made in a reduced, two-dimensional space. To show this, we introduce the change of variables \( p_{\alpha} = R_{\alpha} \omega_{\alpha} \), where the matrix \( R_{\alpha} \) is given as

\[ R_{\alpha} = \begin{bmatrix} p_{x,\alpha} & -p_{z,\alpha} & -p_{x,\alpha} p_{y,\alpha}^2 - p_{y,\alpha} p_{z,\alpha}^2 \frac{p_x}{p_z} & -p_{z,\alpha} p_{x,\alpha} p_{y,\alpha} p_{z,\alpha}^2 \frac{p_z}{p_x} + p_x p_z^2 \\ p_{y,\alpha} & 0 & p_{z,\alpha} p_{y,\alpha} p_{z,\alpha} \frac{p_y}{p_z} & p_{x,\alpha} p_{y,\alpha} p_{z,\alpha} \frac{p_z}{p_y} + p_y p_z^2 \\ p_{z,\alpha} p_{x,\alpha} p_{y,\alpha} & p_{x,\alpha} & p_{x,\alpha} p_{y,\alpha} p_{z,\alpha} p_{y,\alpha} \frac{p_x}{p_z} & p_{x,\alpha} p_{y,\alpha} p_{z,\alpha} \frac{p_z}{p_y} + p_y p_z^2 \end{bmatrix}. \]  

(A6)

One can check that the determinant of \( R_{\alpha} \) is given by \( \text{det}(R_{\alpha}) = -p_{x,\alpha} |p_{\alpha}|^2 \). Introducing the new vector

\[ \rho_{\alpha} = R_{\alpha}^{-1} [f \Delta p_{\alpha} - h_{\alpha}], \]  

(A7)

we have:

\[ \mathcal{P}([r, p] = \prod_{\alpha} \delta(\Delta r_{\alpha} - f^{-1} v_{\alpha}) \mathcal{I}([r, p]), \]  

(A8)

where

\[ \mathcal{I}([r, p]) = \int \prod_{\alpha} d^3 \omega_{\alpha} \exp \left( -\frac{1}{f} \sum_{\alpha} \left\{ i \rho_{\alpha}^T \omega_{\alpha} + D_R \omega_{\alpha}^T \Gamma_{\alpha} \omega_{\alpha} - f \ln (|p_{x,\alpha}|^2) \right\} \right). \]  

(A9)
and the matrix \( \Gamma_\alpha \) has the following form:

\[
\Gamma_\alpha = \begin{bmatrix}
0 & 0 & 0 \\
0 & (p_{x,\alpha}^2 + p_{z,\alpha}^2)|\mathbf{p}_\alpha|^2 & 0 \\
0 & 0 & \frac{p_{z,\alpha}^2|\mathbf{p}_\alpha|^2}{p_{x,\alpha}^2 + p_{z,\alpha}^2}
\end{bmatrix}.
\]

(A10)

The integrals over \( \{\omega_\alpha\} \) in (A9) can now be immediately performed to yield

\[
I[\mathbf{r}, \mathbf{p}] = \left( \frac{\pi f}{D_R} \right)^N \prod_\alpha \left[ \frac{1}{|\mathbf{p}_\alpha|} \delta(f^{-1} \rho_{x,\alpha}) \right] \exp \left( - \frac{1}{4D_R f} \sum_\alpha \left\{ \frac{\rho_{y,\alpha}^2}{(p_{x,\alpha}^2 + p_{z,\alpha}^2)|\mathbf{p}_\alpha|^2} + \frac{\rho_{z,\alpha}^2(p_{x,\alpha}^2 + p_{z,\alpha}^2)}{p_{x,\alpha}^2|\mathbf{p}_\alpha|^4} \right\} \right). \tag{A11}
\]

This result can be written in a more illustrative way if one calculates \( \rho_{x,\alpha} \) explicitly using (A7). One has \( f^{-1} \rho_{x,\alpha} = f^{-1} \rho_{x,\alpha}^f \cdot |f\Delta \mathbf{p}_\alpha - \mathbf{h}_\alpha| \equiv \mathbf{p}_\alpha \cdot \Delta \mathbf{p}_\alpha + 2D_R f^{-1}|\mathbf{p}_\alpha|^2 \). With all this we finally obtain that the probability of a given sequence \( \{\mathbf{r}, \mathbf{p}\} \), in this approximation, is

\[
P[\{\mathbf{r}, \mathbf{p}\}] = \left( \frac{\pi f}{D_R} \right)^N \Sigma[\{\mathbf{r}, \mathbf{p}\}] \exp \left( - \frac{1}{4D_R f} \sum_\alpha \left\{ \frac{\rho_{y,\alpha}^2}{(p_{x,\alpha}^2 + p_{z,\alpha}^2)|\mathbf{p}_\alpha|^2} + \frac{\rho_{z,\alpha}^2(p_{x,\alpha}^2 + p_{z,\alpha}^2)}{p_{x,\alpha}^2|\mathbf{p}_\alpha|^4} + 4fD_R \ln|\mathbf{p}_\alpha| \right\} \right), \tag{A12}
\]

where \( \Sigma[\{\mathbf{r}, \mathbf{p}\}] \) is a singular measure enforcing a set of constraints on the trajectories generated by Eqs. (1) and (2):

\[
\Sigma[\{\mathbf{r}, \mathbf{p}\}] = \prod_\alpha \delta(\Delta \mathbf{r}_\alpha - f^{-1} \mathbf{v}_\alpha) \prod_\alpha \delta(\mathbf{p}_\alpha \cdot \Delta \mathbf{p}_\alpha + 2D_R f^{-1}|\mathbf{p}_\alpha|^2). \tag{A13}
\]

Let us analyze these constraints in more detail. The condition \( \mathbf{p}_\alpha \cdot \Delta \mathbf{p}_\alpha + 2D_R f^{-1}|\mathbf{p}_\alpha|^2 = 0 \) is related the conservation of the norm of \( |\mathbf{p}| \). Note that, in particular, when \( f \to \infty \), it takes the form \( \mathbf{p}(t) \cdot d\mathbf{p}(t) = 0 \), as one would expect. On the other hand, the condition \( \Delta \mathbf{r}_\alpha - f^{-1} \mathbf{v}_\alpha = 0 \) provides, in practice, a practical way to determine \( \rho_{x,\alpha} \) at each time step. Indeed, the tracking device measures the trajectory of the particle, while \( \mathbf{p} \) has to be determined indirectly. Writing this condition more explicitly as \( \Delta \mathbf{r}_\alpha - f^{-1}(\mathbf{v}_\alpha + \mathbf{v}_\alpha^f) = 0 \), one has \( \mathbf{p}_\alpha = v_0^{-1}(f\Delta \mathbf{r}_\alpha - \mathbf{v}_\alpha^f) \), which is the expression that is used in practice to determine \( \rho_{x,\alpha} \) at each time step. In summary, if we restrict our analysis to the relevant sub-space of the trajectories which are compatible with the constraints \( \Sigma \) (forcing, in particular, that \( |\mathbf{p}_\alpha| = 1 \) for all \( \alpha \)), we have the following expression for the discretized path probability:

\[
P_\Sigma[\{\mathbf{r}, \mathbf{p}\}] = \left( \frac{\pi f}{D_R} \right)^N \exp \left( - \frac{1}{4D_R f} \sum_\alpha \left\{ \frac{\rho_{y,\alpha}^2}{(p_{x,\alpha}^2 + p_{z,\alpha}^2)|\mathbf{p}_\alpha|^2} + \frac{\rho_{z,\alpha}^2(p_{x,\alpha}^2 + p_{z,\alpha}^2)}{p_{x,\alpha}^2|\mathbf{p}_\alpha|^4} \right\} \right). \tag{A14}
\]

Appendix B: Experimental set-up and protocol

Bacteria used in this work are smooth swimmer mutants of an \textit{E. Coli} (strain CR20, \( \Delta\text{-CheY} \)) that almost never tumble and were transformed with a plasmid coding for a yellow fluorescent protein (YFP). Bacteria are grown overnight at 30$^\circ$C until the early stationary phase. The growth medium is then removed by centrifuging the culture and removing the supernatant. The bacteria are resuspended in a Motility Buffer (MB) below the very low concentration of \( 3 \times 10^9 \) bacteria per mL, in order to visualize one bacterium at a time and to minimize the interactions between bacteria. The suspension is supplemented with L-serine at 0.08g/mL and polyvinyl pyrrolidone (PVP) at 0.005%; L-serine maintains good motility for a few hours and PVP is used to prevent bacteria from sticking to surfaces. The solution is also mixed with Percoll (1:1) to avoid bacterial sedimentation. The experiments are performed at a temperature of 25$^\circ$C.

The channel is visualized using a home-made Lagrangian tracking microscope [33] here used to track fluorescent swimming bacteria. By a visualization based feedback acting on a mechanical (horizontal) and piezoelectric (vertical) stage, the targeted object is kept close to the center of the visualization field and in focus on an inverted microscope (Zeiss-Observer, Z1 with an objective C-Apochromat 63x/1.2 W). Images of the tracked objects are acquired at 100 Hz with a Hamamatsu Orcaflash 4.0 camera. Simultaneously, the three-dimensional positions of the object are recorded.

The measurements take place in a microfluidic channel of rectangular cross-section (height \( h = 100 \mu m \), width \( W = 600 \mu m \)), made in Polydimethylsiloxane (PDMS) using standard soft-lithography techniques. Flow is imposed through the channel via a Nemesys syringe pump (dosing unit Low Pressure Syringe Pump neMESYS
The flow rate varies from 1 to 4.3 nL s$^{-1}$, corresponding respectively to $U_m$ between (28 ± 1.9) µm/s and (120 ± 4.0) µm/s and maximal shear rates $\dot{\gamma}_M = 4U_m/h$ between (1.12 ± 0.076) s$^{-1}$ and (4.82 ± 0.16) s$^{-1}$. We set our region of interest in the center of the channel with respect to its width and consider only trajectories which are at least 200 µm away from the lateral walls.

...