

**STATISTICAL ANALYSIS OF SELF AND PAIRWISE INTERACTIONS IN  
ACTIVE SYSTEMS**

Katherine M. Daftari

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in  
partial fulfillment of the requirements for the degree of Doctor of Philosophy in the  
Department of Mathematics.

Chapel Hill  
2023

Approved by:

Katherine Newhall

M. Gregory Forest

Pedro Saenz

Kevin Pilkiewicz

Daphne Klotso

©2023  
Katherine M. Daftari  
ALL RIGHTS RESERVED

## ABSTRACT

Katherine M. Daftari: Statistical Analysis of Self and Pairwise Interactions in Active Systems  
(Under the direction of Katherine Newhall)

Active systems often demonstrate impressive group level behaviors which appear coordinated, but are believed to arise only from individual-level interactions. To study the mathematics of these behavioral rules, we reduce such large group dynamics to the simplest cases of self interaction and pairwise interactions. In particular, we will only study passively gathered path data which is recorded without disrupting a system and therefore avoids introducing possible behavior influencing factors. We analyze emergent behaviors in two systems: transient self-trapping in a model of self-avoidant swimming droplets and leader-follower dynamics in experimental golden shiner duos.

We compute several traditional path data analysis metrics including the mean square displacement and two point correlation function of both positional and velocity data to find that they are insufficient to describe the observed dynamics. To address this gap, we propose a use case for estimating the time delayed mutual information of random variables derived from path data; we argue that confined systems are a good candidate for our method since they are likely to reach a steady state. We adapt the sampling scheme of a k-nearest neighbors mutual information estimation method to compute the time delayed self mutual information of the path curvature of self-avoidant swimming droplets to study their self-interactions. We then use the same protocol to estimate the time delayed pairwise mutual information between angular positions of experimental golden shiner duos to study their pairwise interactions.

We find that the decay of the mutual information of the path curvature of self-avoidant swimming droplets can be differentiated from other memoryless models with high path curvature. We also show that the decay timescale of the mutual information relates to the strength of the self-avoidant memory response of these droplets. In our experimental golden shiner path data, we find that peaks in the mutual information curves of the angular positions recover the reaction time or signaling timescale between fish when they are in a leader-follower configuration. Our method is entirely non-parametric and therefore very versatile; it should

be applicable to any path-derived time series data that is either spatially confined or can be shown to sample from a stationary distribution.

A.M.D.G.

## **ACKNOWLEDGEMENTS**

I am grateful to each of my committee members for their guidance and support throughout graduate school. Conversations with Greg were an important factor in my decision to commit to graduate school and I am grateful to Greg Forest and Katie Newhall for giving me early opportunities to be involved in research. Besides being my primary advisor, Katie spent many hours during my early years helping me gain my footing on the first project that I could really call my own. Since then, she has taught me to maintain scientific curiosity during the many evolutions of our project. I also thank Kevin Pilkiewicz and Michael Mayo for their mentorship during our work together and for their continued support as I integrated our work into my dissertation. Thank you to Daphne Klotsa and Pedro Saenz for their feedback which helped me understand the broader implications of my research.

I am thankful to all of my friends in the department and outside of it for celebrating my wins with me and helping me persevere through challenges with their companionship. The support of my parents and my sisters and brothers has been a comfort during graduate school and their examples of kindness, virtue, and hard work have shaped who I am and who I hope to be. Most of all, I am moved by the unconditional support and patience that Kamron has shown me over the last few years. His unwavering belief in me at all times is the reason that I have made it to this finish line, and I am excited to see which new finish lines we will cross together.

## TABLE OF CONTENTS

LIST OF TABLES .....	x
LIST OF FIGURES .....	xi
LIST OF ABBREVIATIONS .....	xv
1 Introduction .....	1
1.1 Active Systems.....	1
1.2 Key Goals .....	2
2 Mathematical Methods: Information Theory .....	4
2.1 Information Theory .....	4
2.1.1 Mutual Information.....	5
2.1.2 Estimating Mutual Information: KNN Methods .....	6
2.1.3 Mutual Information Examples.....	9
2.1.4 Some Properties of Mutual Information.....	10
2.2 Time Delayed Mutual Information of an Ensemble of Independent Agents Evolving in Time .....	11
2.3 Adaptations for a Single-Path Estimator .....	13
2.3.1 Limitations and Advantages.....	17
3 Exploration of a Model for Swimming Droplets .....	19
3.1 Introduction.....	19
3.2 A Model for Self-avoidant Memory .....	21
3.3 Comparative Analysis .....	25
3.3.1 Intermediate Time Scales: Ballistic Motion .....	26
3.3.2 Long Time Scales: Enhanced Diffusion .....	30
3.4 Limiting Behavior.....	32

3.5	Conclusions .....	35
4	Unpacking Emergent Behaviors of Self-Avoidant Swimming Droplets with Curvature Statistics ...	37
4.1	Introduction.....	37
4.2	Calculating Curvature Directly .....	38
4.3	Using Mutual Information to Estimate Nonlinear Correlations in Path Curvature .....	41
4.4	Time Delayed Self Mutual Information of a Single Agent .....	46
4.5	Individual Path Analysis .....	47
4.6	Tuning Memory .....	50
4.7	Comparison Across Models.....	55
4.8	Conclusions.....	57
4.9	Future Work .....	59
5	Using Trajectory Data to Infer the Character of Social Interactions Between Golden Shiners .....	60
5.1	Introduction.....	60
5.2	Experimental Methods .....	61
5.3	Data Preparation and Examination .....	65
5.4	Mutual Information Setup and Variable Selection .....	66
5.5	Reduced Model .....	69
5.6	Experimental Data Analysis .....	77
5.7	Experimental Summary Analysis.....	79
5.8	Second Model .....	80
5.9	Conclusions.....	81
5.10	Future Work .....	82
6	Conclusions .....	90
	APPENDIX A: SAMPLING SCHEME AND MUTUAL INFORMATION ESTIMATION CODE.....	93
A.1	Time Delayed Sampling Scheme .....	93
A.2	KNN Estimator for Mutual Information .....	95
	APPENDIX B: ADDITIONAL CALCULATIONS FOR SWIMMING DROPLET MODEL .....	98
B.1	Solving the Diffusion Equation to Combine and Nondimensionalize the Coupled System ....	98

B.2	Computation of the Velocity Integral Formulation Using a Dirac Delta Function .....	99
B.3	Computation of the Hover Height Integral Formulation.....	101
B.4	Computing the Small Time Asymptotics of the Active Brownian MSD.....	103
B.5	Computing MSD and OCF from Position Time Series Generated by the Model .....	104
APPENDIX C: SUPPLEMENTAL FIGURES TO GOLDEN SHINER BEHAVIORAL ANALYSIS...		106
C.1	Supplemental Figures to Experimental Golden Shiner Behavioral Analysis .....	106
BIBLIOGRAPHY .....		116

## LIST OF TABLES

5.1	Experimentally estimated average lap times. ....	80
-----	--	----

## LIST OF FIGURES

2.1	Entropy of a coin toss .....	5
2.2	Anscombe's Quartet .....	7
2.3	Nearest neighbor schematic .....	8
2.4	Similar turning times generate different curvature features in different models .....	9
2.5	Mutual information is invariant to transformations .....	10
2.6	Sample trajectories of Ornstein-Uhlenbeck walkers .....	12
2.7	Time delayed samples of Ornstein-Uhlenbeck walkers .....	12
2.8	Mutual information of ensemble of Ornstein-Uhlenbeck walkers .....	13
2.9	Mutual information of a Gaussian walker .....	15
2.10	Average separation window $W$ between consecutive samples .....	16
2.11	Distribution of separation windows .....	16
2.12	Time-delayed sampling scheme for mutual information calculation .....	17
3.1	Sample paths of self-avoidant model .....	27
3.2	Statistics of self-avoidant model (MSD, OCF) .....	29
3.3	Parameter space of self-avoidant model .....	31
3.4	Transient self-caging in self-avoidant model paths .....	33
4.1	Self-trapping is seen in self-avoidant model paths .....	39
4.2	MSSI Sampling Scheme .....	40
4.3	Average MSSI values for self-avoidant model paths .....	40
4.4	Using MSSI to find regions of self-trapping .....	42
4.5	MSSI ensemble histogram and empirical CDF over all times .....	43
4.6	Ensemble time evolving mean and variance .....	44
4.7	Ensemble time-delayed sampling scheme .....	45
4.8	Ensemble time-delayed mutual information evolution .....	46
4.9	Time-delayed sampling scheme for mutual information calculation .....	47

4.10	Autocorrelation of path straightness .....	47
4.11	Mutual information decay comparison between paths .....	49
4.12	Mutual information decay fits of various paths .....	50
4.13	Illustrated effects of $\mu$ and $M$ on concentration field .....	53
4.14	Changing velocity $V$ with effective memory window $M$ .....	54
4.15	Changing path dynamics with effective memory window $M$ .....	54
4.16	Ensemble average MI decay timescale vs memory integral truncation .....	55
4.17	MI decay comparison across models .....	57
5.1	Live golden shiner .....	61
5.2	Experimental setup aerial photo .....	63
5.3	Four main classes of behavior are seen in experimental data .....	64
5.4	Alignment angle schematic .....	67
5.5	Alignment angle data .....	67
5.6	Alignment angle data .....	68
5.7	Alignment angle data .....	68
5.8	Alignment angle data .....	69
5.9	Mutual information matrix of variable pairs.....	70
5.10	Sample toy model trajectories .....	71
5.11	Toy model mutual information decay curves .....	74
5.12	Toy model mutual information decay curves .....	74
5.13	Toy model correlation .....	75
5.14	Toy model correlation .....	75
5.15	Model average lap time .....	76
5.16	Toy model mutual information at $ T  = T^*$ .....	83
5.17	Toy model correlations at $ T  = T^*$ .....	83
5.18	Experimental mutual information decay .....	84
5.19	Experimental correlation .....	84

5.20	Experimental mutual information decay .....	85
5.21	Experimental mutual information LOESS fit varies with $f$ .....	85
5.22	Experimental mutual information peak location varies with $f$ and $W$ .....	86
5.23	Mutual information decay of all experiments .....	86
5.24	Correlation decay of all experiments .....	87
5.25	Model mutual information decay with increased follower noise .....	88
5.26	Model mutual information decay with increased follower noise .....	88
5.27	Model mutual information decay with increased follower noise .....	89
B.1	Droplet hover height .....	102
B.2	Sampling schematic for non-overlapping windows in MSD, OCF calculation .....	105
C.1	AS-1 Alignment Data .....	106
C.2	AS-2 Alignment Data .....	106
C.3	AS-3 Alignment Data .....	107
C.4	AS-4 Alignment Data .....	107
C.5	AS-5 Alignment Data .....	108
C.6	CS-2 Alignment Data .....	108
C.7	CS-3 Alignment Data .....	109
C.8	CS-4 Alignment Data .....	109
C.9	CS-5 Alignment Data .....	110
C.10	AS-1 MI Decay .....	110
C.11	AS-2 MI Decay .....	111
C.12	AS-3 MI Decay .....	111
C.13	AS-4 MI Decay .....	111
C.14	AS-5 MI Decay .....	112
C.15	CS-2 MI Decay .....	112
C.16	CS-3 MI Decay .....	113
C.17	CS-3 MI Decay .....	113

C.18 CS-5 MI Decay .....	113
C.19 Experimental mutual information peak location varies with $f$ .....	114
C.20 Experimental mutual information peak location varies with $f$ .....	115

## LIST OF ABBREVIATIONS

ABP	Active Brownian Particle
AS	Agitated State
BEL	Behavioral Engineering Lab
CS	Control State
ERDC	Engineer Research and Development Center
KNN	K-Nearest Neighbors
LOESS	Locally Estimated Scatterplot Smoothing
MI	Mutual Information
MIPS	Motility Induced Phase Separation
MSD	Mean Square Displacement
MSSI	Multi-scale Straightness Index
OCF	Orientation Correlation Function
PDE	Partial Differential Equation
SDE	Stochastic Differential Equation
SDS	Sodium Dodecyl Sulfate
VCF	Velocity Correlation Function

# CHAPTER 1

## Introduction

### 1.1 Active Systems

Active systems are comprised of one or many individual living or nonliving units that consume or use energy to produce mechanical work used for locomotion (movement). A particular feature of interest within active systems of all types and scales is the emergence of collective behavior [77]. Such large and seemingly coordinated group level behavior is a little understood phenomenon in which many independent individual interactions create the appearance of intelligent group behavior. Shoaling [24, 43, 69], swarming [52, 9, 8], flocking [4, 12], and herding [22] are all examples of emergent behavior in macroscale living systems which have been studied extensively. These behaviors are believed to serve evolutionary purposes such as protection from predators [43, 22] and more efficient foraging [22]. On the microscale, many types of bacteria have been shown to respond to different environmental stimuli, such as local gradients in chemical concentrations, [1, 73, 2], gravity [26, 13], or light [66, 29, 79].

New data gathering methodologies and computational methods have made active systems a rich and continually evolving scientific field with many inventive synthetic/nonliving systems being inspired by these biological active systems. When such systems exist at a microscale, we refer them as active particles, which derive their motility from environmental energy consumption that is transformed into self-propulsion. These self-propelled particles include (but are not limited to) autophoretic swimming droplets [74, 44, 35, 33, 30], chemically propelled droplets [54, 36, 72, 34], and even light sensitive particles [61, 11, 50, 49]. (For a comprehensive review of micro-scale active systems and current research developments, see Refs. [80, 17, 18].) Non-microscopic systems of autonomous robots or hexbugs have also been studied [28, 81, 16]. Across critical parameter thresholds, these systems will often exhibit motility induced phase separation (MIPS), wherein clusters of agents form a “solid-like” state surrounded by a “fluid-like” bath of motile individual agents [3, 21, 71]. The regimes in which this behavior appears and the real forces or effective forces which maintain such clusters is a topic of great interest within the soft matter community [37, 20, 19, 23].

In all cases, the common theme of such emergent behaviors is the absence of any group-level coordination or leadership. Exploration of this phenomenon has led to the development of computational models that can be categorized into two main classes: continuum modeling and agent based modeling. By definition, active systems are out-of-equilibrium since each agent continually dissipates energy into the environment, and the system breaks time reversal symmetry. Therefore, the system is not relaxing towards an equilibrium state. Despite this, continuum models have developed to explore the dynamics of the density of the ensemble [67, 65, 59], and in certain cases, can be treated as an ideal gas [23]. In contrast, agent based models prescribe mathematical rules by which each individual agent evolves in time. Often the model will include “interaction rules” in which the position and state of an agent is informed by the position and state of surrounding agents and/or evolving ensemble parameters [76]. Such deterministic rules are usually implemented with a stochastic component (usually white noise) which mimics the observed apparent randomness often seen in nature.

## 1.2 Key Goals

Despite their importance, the exact behavioral rules of many active systems (especially living ones) at the individual level are still poorly understood. To pursue this question, we reduce the problem to the most fundamental case and investigate self and pairwise interactions using an agent-based modeling approach. Specifically, we aim to characterize the physical pathways or mechanisms by which agents in two distinct systems sense and respond to stimuli. In particular, we examine the temporal features of path data, which is a common experimentally gathered data type since it can be passively gathered without disrupting the system.

The first system which we study is a theoretical model of microscale swimming droplets. This model is physically inspired by experimental droplets which “swim” in a surfactant bath due to local surface tension instabilities which create tiny microcurrents (Marangoni flows) that propel the droplets in a ballistic fashion [74, 44, 35, 33, 30]. Such directional swimming droplets can be fabricated or “programmed” to prefer different environmental chemical regimes. In particular, our droplets navigate a changing chemical environment by seeking out the regions of lowest chemical density, and in doing so, they avoid moving toward locations where they have already been. For this reason, we label them “self-avoidant.” (In biology, this type of environment-mediated interaction is called stigmergy and has been seen in social insects such as ants and termites [70].) As they move, the droplets sometimes trap themselves in their own chemical trails and create unique spiral-like path features that we believe are an *emergent effect* of the memory. This path

feature is not explicitly determined by model equations, rather, the spontaneous ordering in the curvature arises unpredictably as a result of interactions at the smallest scale (thermal noise). We cannot predict which paths will exhibit this feature, rather, we can merely let the system evolve and observe the outcome.

The second system we study is a biological living system of confined golden shiner pairs, which are social organisms that are used frequently in behavioral experiments. In our experimental video data, we observe that our golden shiner pairs appear to cycle through similar modes of behavior *together*, which is highly suggestive of interaction or coordination between fishes. However, throughout the experiment, all environmental factors remain constant, which suggests that the fish transition in and out of these apparent coordinated motion states spontaneously. (The transitions are not preempted by any external stimuli.)

In Chapter 3, we develop an agent based model for a self-avoidant swimming droplet. We thoroughly explore the path data using conventional tools and we find a parameter regime in which the self-avoidant memory response causes transient self-trapping. We show that a commonly used memoryless model (active Brownian particle model) cannot explain the exotic path features seen in the self-avoidant model. In Chapter 4, we use our novel approach for measuring nonlinear correlations in time series data, and show that our approach can quantify different levels of memory expression between individual paths, and between ensembles with average memory strength differences. In Chapter 5, we continue our novel approach and apply it to experimental data of a living system and find that our results are in good agreement with a contrived model. Using our metrics, we provide strong evidence that the signalling timescale between two individual fishes can be non-parametrically estimated from only the path data. Our approach in both Chapters 4 and 5 will rely heavily on tools from information theory, specifically mutual information, which we introduce in Chapter 2. Following this introduction, we present our mathematically rigorous methods for using a mutual information estimator on time series data. In our methods we identify appropriate conditions for utilizing mutual information on time series data and derive our method with careful treatment of ambient dynamical correlations in time series data, which are overlooked in current literature.

## CHAPTER 2

### Mathematical Methods: Information Theory

#### 2.1 Information Theory

The foundations of modern information theory were developed in the early 20th century; the most notable of the early contributors was Claude Shannon after whom the foundational concept of Shannon entropy is named. In its infancy information theory was created to study and improve the transmission of digital information; the primary applications were the telegraph and cryptography. Since then, information theory has expanded in its use. The information, or self-information of a random variable was introduced by Claude Shannon as an alternative way of expressing the probability of a given event (like the odds of an event, for example). The information, or *surprisal*, of a discrete random variable  $X$  with probability mass function  $p(x)$  maps each possible event outcome to an associated information content [63]:

$$I(X) = \log \frac{1}{p(x)}. \quad (2.1)$$

High probability events have low information (they are unsurprising), and low probability events have high information- a rare event will require a lot of information to encode. To communicate the average amount of surprise in a given random variable  $X$  with probability mass function  $p(x)$ , we use the Shannon entropy:

$$H(X) = - \sum_{x \in X} p(x) \log(p(x)). \quad (2.2)$$

The base of the logarithm determines the units of  $H(X)$ ; base 2 has units of bits, base  $e$  uses units of nats. Entropy is deeply related to information- the more information is needed to encode an outcome, the more uncertainty or entropy the outcome contains. In [60], the Shannon entropy is characterized as the optimal amount of information needed to encode independent draws from the random variable  $X$ .

As an example, consider a Bernoulli trial, or a coin toss, where  $\theta$  measures the probability of heads as the outcome and the probability of tails is  $1 - \theta$ . Then,

$$H(\theta) = - \sum_X p(x) \log(p(x)) = - [(1 - \theta) \log(1 - \theta) + \theta \log(\theta)],$$

which is maximized when  $\theta = \frac{1}{2}$ , also shown in Fig. 2.1. Thus, we have maximal entropy or uncertainty when we have a fair coin, or when  $\theta = \frac{1}{2}$ , and minimal predictive capability at the same place. Continuing the coin toss example, the value of  $H(\theta = \frac{1}{2}) = -\log_2(\frac{1}{2}) = \log_2(2) = 1$ , meaning that it takes 1 bit (in this example we use base 2 for the logarithm) of information to represent the outcome of the coin toss where  $\theta = \frac{1}{2}$ . Thus, we need maximal information to communicate the outcome of an experiment where there is maximum entropy. When there is higher certainty about the outcome, such as when  $\theta = \frac{3}{4}$ , there is correspondingly lower entropy ( $H(\theta = \frac{3}{4}) < 1$ ), because we need less information to “predict” the result.

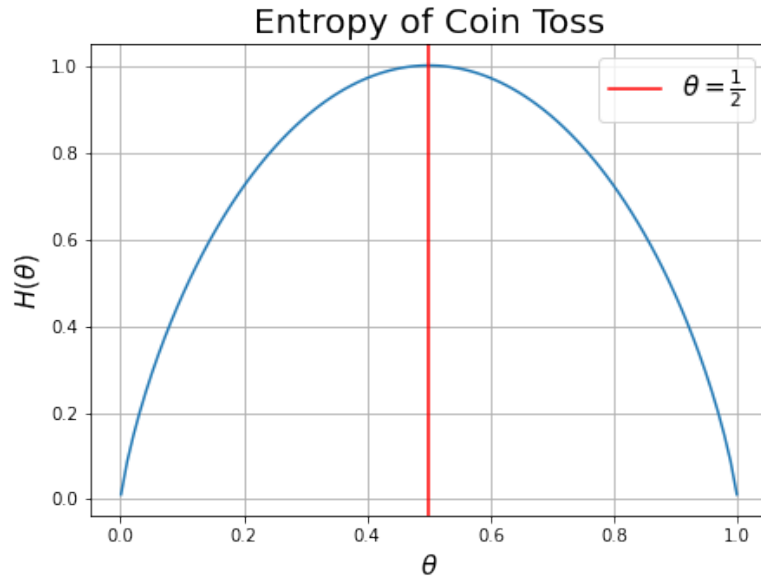


Figure 2.1: For a single coin toss with probability  $\theta$  of heads, the entropy is maximized in a fair coin when  $\theta = \frac{1}{2}$ .

### 2.1.1 Mutual Information

Ultimately, the self-information and Shannon entropy tell us about the characteristics of a random variable by itself, and we are interested in the relationships between two random variables. If a linear relationship can

be assumed, we might use correlation to determine the relationship between two variables. Often, a linear relationship cannot be assumed, therefore, we want to assess the strength of nonlinear correlations between two random variables. Shannon offered a solution to this problem in [63], which was later called the *mutual information* between  $X \sim p_X(x)$  and  $Y \sim p_Y(y)$  with supports  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively:

$$MI(X; Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} dx dy. \quad (2.3)$$

$MI(X; Y)$  is formally the Kullback-Liebler divergence between the product of the marginal distributions,  $p_X(x) \cdot p_Y(y)$  and the joint distribution,  $p_{X,Y}(x, y)$ . Intuitively, we can interpret the result as the “distance from independence” because the integral is formally zero in the case where  $X$  and  $Y$  satisfy ( $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ ) and is infinite in the case where  $p_X(x) = p_Y(y)$ . Mutual information of  $O(1)$  between samples is typically interpreted as a strong signal.

We illustrate the importance of including nonlinear correlations using Anscombe’s Quartet, shown in Fig. 2.2. This unique dataset, constructed by statistician Francis Anscombe, has nearly identical summary statistics: mean, variance, correlation, and even linear fit up to 3 decimal places. However, the data shapes are all quite different. We find that the mutual information, which measures all nonlinear correlations, does a better job of distinguishing the three datasets. Dataset  $(X1; Y1)$ , which has a clearly nonlinear (parabolic-like) shape has the highest mutual information, and dataset  $(X3; Y3)$  has the lowest information. The low value of  $MI(X3; Y3)$  indicates that knowing the value of  $X3$  gives us little information about the value of  $Y3$ . We see that Anscombe’s Quartet clearly illustrates why assuming a linear relationship between  $X$  and  $Y$  could misrepresent the true relationship between  $X$  and  $Y$ .

### 2.1.2 Estimating Mutual Information: KNN Methods

To compute the mutual information, we use method 1 detailed in [32], which uses a k-nearest neighbors method to estimate the entropy of the underlying marginal densities of our random variables. The algorithm is summarized in the following paragraph. Begin with a set  $\{z_i\}, i = 1, \dots, N$  of independent, identically distributed bivariate observations of some jointly distributed random variable  $Z(X, Y)$  with joint density  $\mu(x, y)$ . (Note that  $x$  and  $y$  can be multidimensional.) We choose a norm to calculate the kth nearest neighbor to each point,  $z_i$ . Kraskov uses the max norm:  $|z_i - z_j| = \max\{|x_i - x_j|, |y_i - y_j|\}$  where  $z_i$  is the ordered pair  $(x_i, y_i)$  and  $z_j$  is the ordered pair  $(x_j, y_j)$ . To compute the max norm, for each  $(x_i, y_i)$  we select the

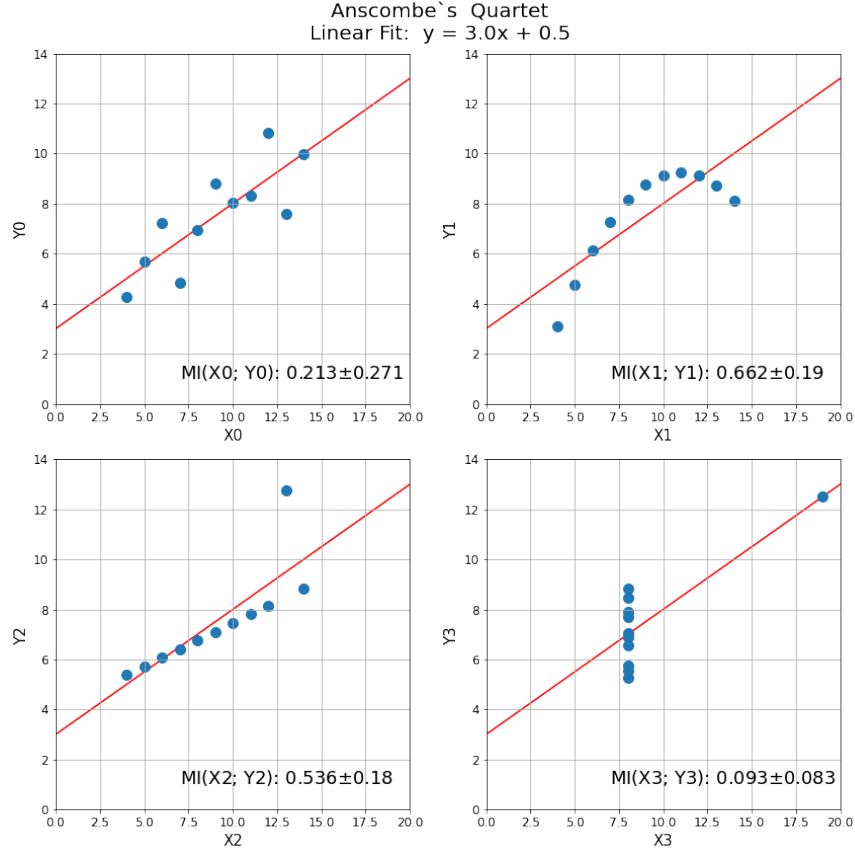


Figure 2.2: An identical linear fit for four very different datasets illustrates why the assumption of a linear relationship between datasets can fall short. We compute the mutual information numerically (using the method detailed later in this section) for each dataset using a sample size of  $N = 6$  and 50 repetitions. Low amplitude white noise (with strength  $O(10^{-4})$ ) was added to  $X3$  data in keeping with the suggestion in [32].

maximum of the  $x$ -subspace and  $y$ -subspace distances from all  $x_j$  and  $y_j$ ,  $j \neq i$  to  $x_i$  and  $y_i$ . These subspace distances are denoted  $\epsilon_x(i)/2$  and  $\epsilon_y(i)/2$ . For selected  $k$ , the max norm corresponding to  $z_i$  is the  $k^{th}$  smallest element of the previous list, which we call  $\epsilon(i)/2$ . (Observe that  $\epsilon(i)$ ,  $\epsilon_x(i)$ , and  $\epsilon_y(i)$  are all random variables.) Following, we count the number of points  $x_j$  ( $j \neq i$ ) where  $|x_i - x_j| < \epsilon(i)/2$ ; call this  $n_x(i)$ . Similarly for  $y$ . Repeat this process for all  $\{z_i\}$  resulting in two values,  $n_x(i)$  and  $n_y(i)$  for each  $z_i$ . Finally, we compute

$$I^{(1)} = \psi^{(0)}(k) - \langle \psi^{(0)}(n_x + 1) + \psi^{(0)}(n_y + 1) \rangle + \psi^{(0)}(N) \quad (2.4)$$

where  $\langle \psi^{(0)}(n_x + 1) + \psi^{(0)}(n_y + 1) \rangle = \frac{1}{N} \sum_{i=0}^N [\psi^{(0)}(n_x(i) + 1) + \psi^{(0)}(n_y(i) + 1)]$  and the digamma function is the first logarithmic derivative of the gamma function:  $\psi^{(0)}(t) = \frac{d}{dt} \ln(\Gamma(t))$ . (It can be approximated as  $\psi^{(0)}(t) \sim \ln(t) - \frac{1}{2t}$ .)

It can be seen from Eq. 2.4 that higher values of  $n_x$  and  $n_y$  corresponding to greater subspace density and therefore lower subspace entropy will decrease the value of  $I$ . Conversely, lower values of  $n_x$  and  $n_y$  will increase the mutual information,  $I$ . This agrees with our intuition that “rare” outcomes which will be associated to regions in the joint density domain with lower probability mass will contain greater information, or they are more surprising.

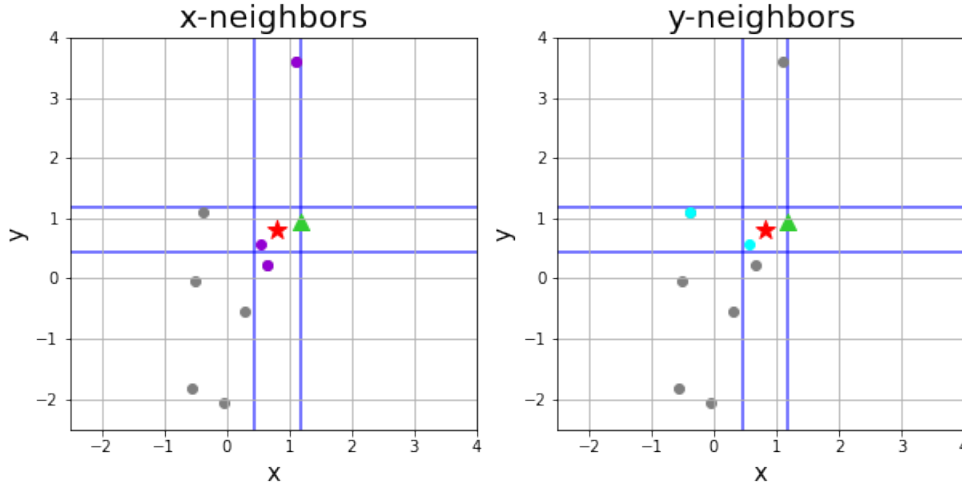


Figure 2.3: Ten draws from a bivariate normal distribution with zero mean and unit variance. For the selected data point  $(x_i, y_i)$  (red star), the nearest neighbor is indicated by a green triangle. A corresponding box centered at  $(x_i, y_i)$  of side length  $\epsilon_i$  is shown in blue. In the x-dimension, four neighbors are found (purple) including the nearest neighbor (green). In the y-dimension, three neighbors are found (cyan) including the nearest neighbor (green).

In Fig. 2.3, we show a random sample of  $N = 10$  points from a bivariate Gaussian with mean zero and unit variance. For the selected point (red star), we find the first nearest neighbor (green triangle), corresponding to  $k = 1$ . The blue lines mark a box centered at the selected point of side length  $\epsilon(i)$  where the distance between the selected point and its nearest neighbor is  $\epsilon(i)/2$ . In the x-dimension, there are three points whose x-coordinate distances are strictly less than  $\epsilon(i)/2$ , therefore  $n_x(i) = 3$ . (Note that we exclude the nearest neighbor since  $\epsilon_x(i)/2 = \epsilon(i) = 2$  in this case. i.e., the x-coordinate distance between points determined the box width.) In the y-dimension, there are three points whose y-coordinate distances are strictly less than  $\epsilon(i)/2$ , therefore  $n_y(i) = 3$ .

### 2.1.3 Mutual Information Examples

In Fig. 2.4 we show how the mutual information of a bivariate Gaussian with correlation coefficient  $\rho$  changes as a function of sample size. The true mutual information between two Gaussian random variables is known to be  $MI(X; Y) = -\frac{1}{2} \log(1 - \rho^2)$  [32], which increases as  $\rho$  increases. (This corresponds to our intuition that two Gaussians with high linear correlation coefficient  $\rho$  will also have high nonlinear correlations and therefore high mutual information.) As the sample size increases we see that the mutual information rapidly converges to the true value and the standard errors decrease, consistent with the same finding presented in [32]. We note that when  $\rho = 0$ , the Gaussians are totally uncorrelated (formally independent) and the true mutual information is exactly zero; we can use the standard errors of the curve corresponding to  $\rho = 0$  as an estimate of the amount of noise expected when using this estimator.

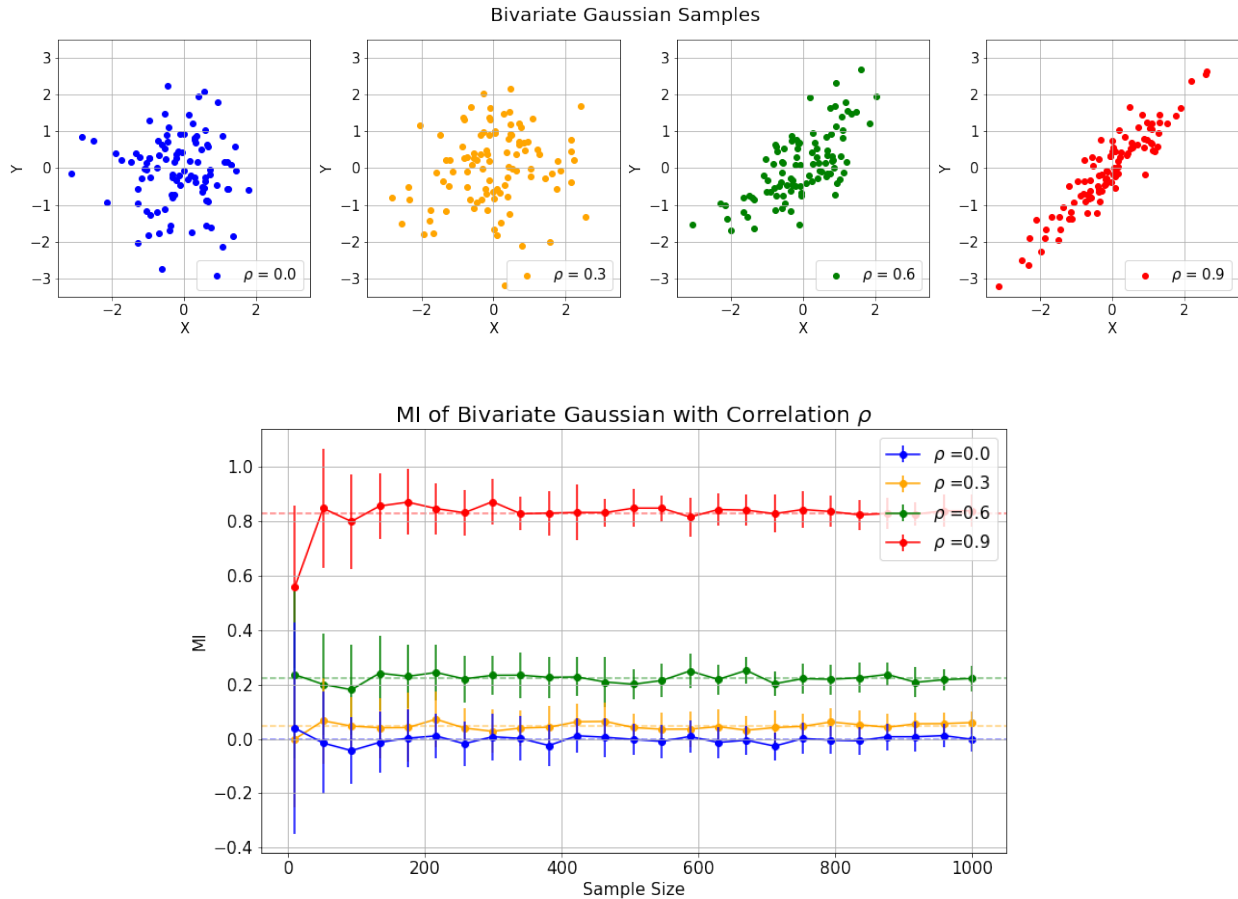


Figure 2.4: (a) Bivariate Gaussian samples with zero mean and unit variance ( $N = 200$ ). (b) As we change the correlation coefficient,  $\rho$ , in a bivariate normal distribution, the mutual information increases, as expected. Each data point shows the average and standard errors of 30 repetitions. Dashed lines indicate the true mutual information value of a bivariate Gaussian with correlation  $\rho$ :  $MI(X; Y) = -\frac{1}{2} \log(1 - \rho^2)$  [32].

### 2.1.4 Some Properties of Mutual Information

One of the most important properties of mutual information is that it is symmetric, therefore  $MI(X; Y) = MI(Y; X)$ . Due to this symmetry, other information theoretic measures, such as transfer entropy [60] and causation entropy [68] are sometimes preferred since they can estimate directional information flow between agents under the correct conditions. However, such directional metrics require conditioning on the past history of an agent, which can be difficult to do with experimental data. In our adaptations for time series data (Section 2.3), we propose that the inclusion of a time delay into the mutual information calculation can provide a statistically sound and computationally efficient workaround to this issue.

Another interesting property of the mutual information between two distributions  $X$  and  $Y$  is that  $MI(X; Y) = MI(X'; Y')$  for  $X' = F(X)$  and  $Y' = G(Y)$  where both  $F$  and  $G$  are smooth and invertible maps [32]. (Note that entropy is not invariant to transformations [32]). We illustrate this fact in Fig. 2.5 where we compute the mutual information between a multivariate Gaussian random variable and the transformation  $F(\cdot) = e^{\mu + \sigma(\cdot)}$ , making  $X' = e^{\mu + \sigma X}$ , which is lognormally distributed. (Similarly for  $Y$ ).

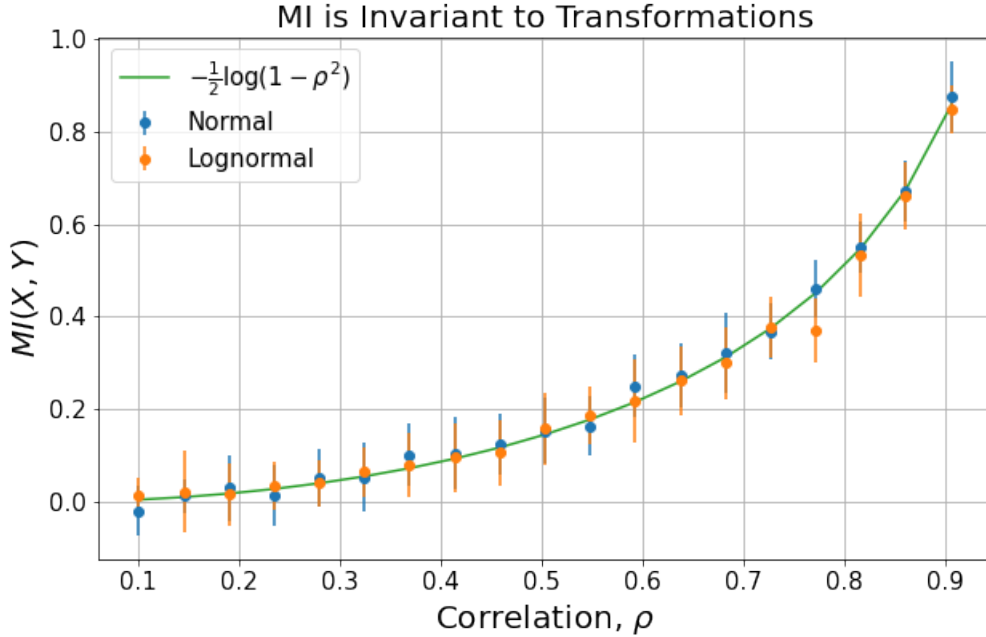


Figure 2.5: For a bivariate Gaussian random variable with correlation coefficient  $\rho$ , the exact mutual information can be computed as  $MI = -\frac{1}{2} \log(1 - \rho^2)$ . We compare the mutual information of a multivariate normally distributed random variable with a lognormally distributed random variable with randomly selected means and changing variance. Consistent with Kraskov, we find that the mutual information does not change.

## 2.2 Time Delayed Mutual Information of an Ensemble of Independent Agents Evolving in Time

To develop the tools for using mutual information to analyze time series data, we will first show how the mutual information can be used to analyze a time evolving ensemble. Here, we show how to estimate the nonlinear correlations between the state of an evolving system of  $N$  agents at time  $t$  and at future time  $t + T$ . First, consider an ensemble of  $N$  identical independent agents whose positions (or any other measurable quantity) evolve according to an Ornstein-Uhlenbeck process. Then, the position of agent  $n$  follows the stochastic differential equation:

$$dX_n(t) = \kappa(\theta - X_n(t))dt + \sigma dW_n(t) \quad (2.5)$$

where  $\kappa$  is the mean reversion coefficient to the asymptotic mean,  $\theta$ . The parameter  $\sigma$  controls the strength of the additive white noise which is modeled as increments of a Wiener process ( $W(t)$ ). The Ornstein-Uhlenbeck process is also known as a mean-reverting Gaussian walk; the particle experiences a tendency to walk back towards the asymptotic mean  $\theta$  and this tendency increases as the distance from  $\theta$  increases. In this way, the walker experiences an effective attracting force toward  $\theta$ . While this is not exactly a confined system since the walker *could* stray very far from  $\theta$  (although with low probability), this mean reversion property has important implications for the ensemble mutual information.

In our ensemble, the  $N$  trajectories are labeled  $\{\{X_1(t)\}, \{X_2(t)\}, \dots, \{X_N(t)\}\}$ , where  $\{X_n(t)\}$  represents a finite time series describing the temporal evolution of the position of agent  $n$  from time  $t = t_0$  to time  $t_F$ . We show sample trajectories of 5 Ornstein-Uhlenbeck walkers in Fig. 2.6. Since each agent is identical and independent from all other agents, it is true that the selection of all agents' positions at any particular time  $t_i \in \{t_0, \dots, t_F\}$ , given by  $X(t_i) = \{X_1(t_i), X_2(t_i), \dots, X_N(t_i)\}$ , is a collection of  $N$  independent samples from the true distribution of positions of the ensemble at time  $t_i$ . Similarly, at any particular point after time  $t_i$ , such as time  $t_i + T$ , with  $T > 0$ , the set  $X(t_i + T) = \{X_1(t_i + T), X_2(t_i + T), \dots, X_N(t_i + T)\}$  is also a collection of  $N$  independent samples from the distribution of positions of the ensemble at time  $t_i + T$ . We illustrate this in Fig. 2.7 for  $t = 35\text{s}$  (magenta) and  $T = 4\text{s}$  (green).

Since we have two independent samples, we can compute the mutual information  $MI(X(t_i); X(t_i + T))$  which measures the strength of all nonlinear correlations between the positions of the ensemble at time  $t_i$  and

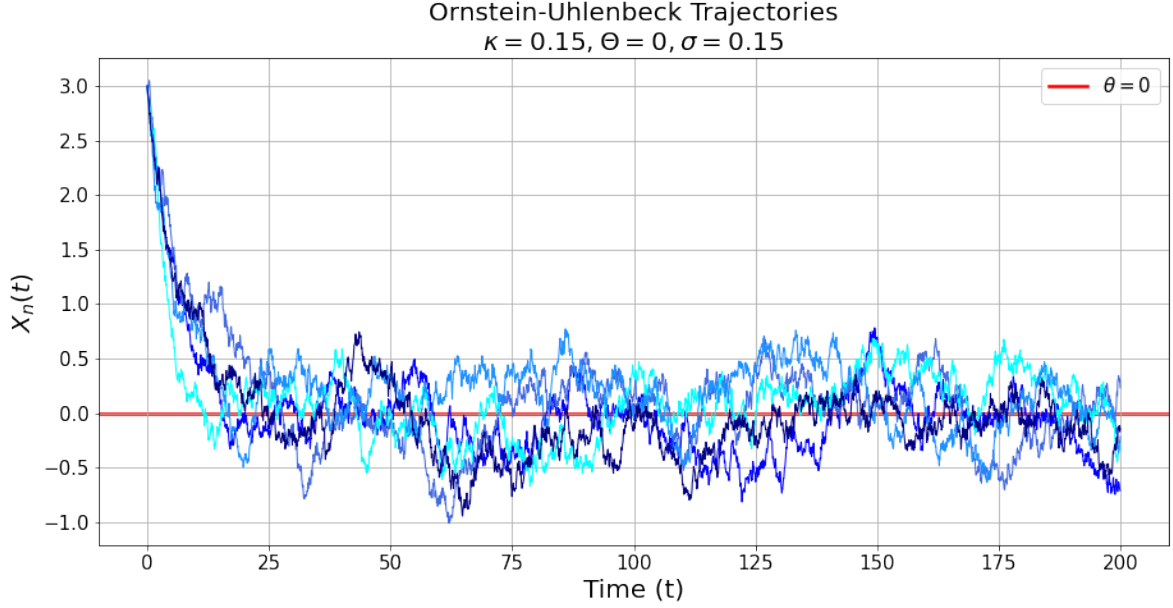


Figure 2.6: Sample trajectories of Ornstein-Uhlenbeck walkers.

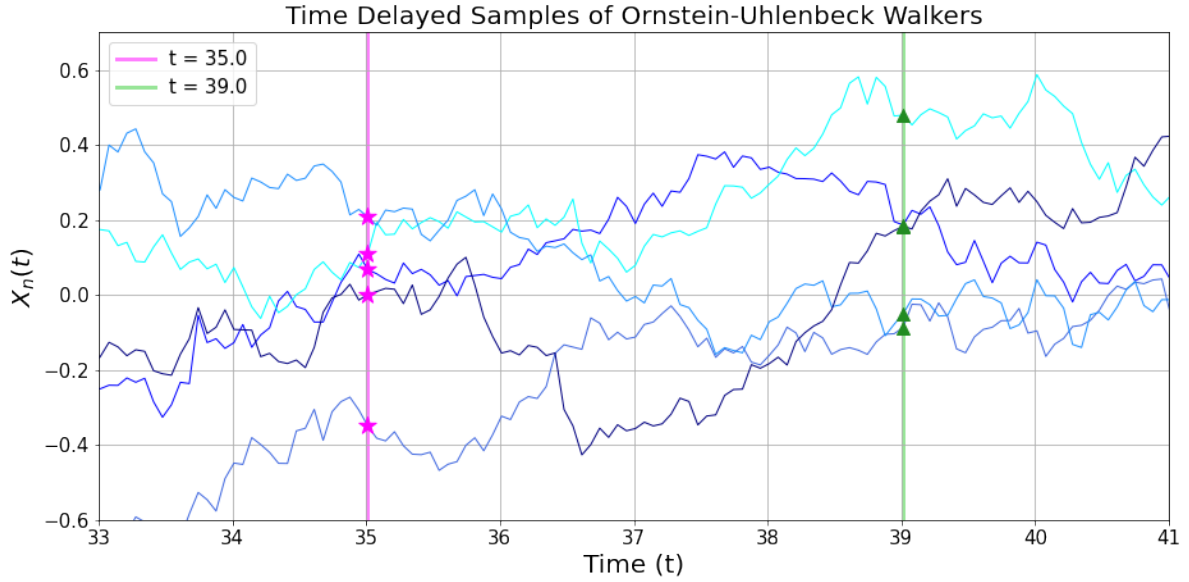


Figure 2.7: Time delayed samples of 5 Ornstein-Uhlenbeck walkers from the ensemble in Fig. 2.6.

time  $t_i + T$ . Since we can compute the mutual information  $MI(X(t_i); X(t_i + T))$  for any  $t_i$ , and in fact all  $t_i \in \{t_0, \dots, t_F\}$  (with  $t_i + T \leq t_F$ ), we can compute the evolution of the mutual information from time  $t_0$  to  $t_F$  for any time delay  $T$ . We show the time evolving mutual information of  $N = 100$  Ornstein-Uhlenbeck walkers for varying time delays  $T$  in Fig. 2.8. At small times  $t$ , we see that the mutual information is initially

small. This is consistent with our trajectory data (sample shown in Fig. 2.6), where all of the walkers start from the same initial position. Therefore, knowing the position of all walkers at time  $t_0$  only reduces the uncertainty of knowing the positions at time  $t_0 + T$  by a small amount, meaning that the shared mutual information is small. At longer times  $t$ , the distribution of positions approaches a steady state and the mutual information correspondingly fluctuates around a constant value for fixed  $T$ . In the case that the ensemble's positions are smoothly evolving in time, we expect then that the distributions of positions at time  $t_i$  and time  $t_i + T$  are similar (for small enough  $T$ ) and therefore that  $MI(X(t_i); X(t_i + T))$  decreases as  $T$  increases. We see this is confirmed in Fig. 2.8; at larger values of  $T$  the mutual information decreases.

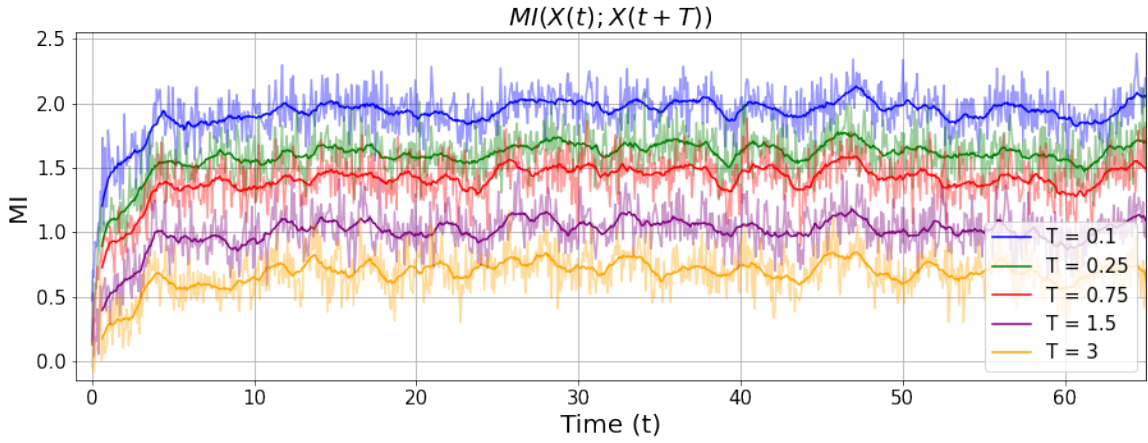


Figure 2.8: Time evolving ensemble mutual information of 100 Ornstein Uhlenbeck walkers.

### 2.3 Adaptations for a Single-Path Estimator

Suppose now that we want to estimate the time delayed mutual information on a smaller scale, between a pair of agents,  $MI(X_n(t); X_m(t + T))$ , or the time delayed self mutual information of a single agent, given by  $MI(X_n(t); X_n(t + T))$ . We will introduce our adaptation using the case of the time delayed mutual information of a single agent, which is the amount of nonlinear correlation between the position of agent  $X_n$  at time  $t_i$  and future time  $t_i + T$ . Consider that both  $X_n(t_i)$  and  $X_n(t_j)$  are realizations of the continuous stochastic process  $X_n$ , and therefore we expect that  $X_n(t_i)$  and  $X_n(t_j)$  will be correlated to some extent, especially if  $t_i$  and  $t_j$  are very close in time. (The same is true of  $X_n(t_i + T)$  and  $X_n(t_j + T)$ ). We are challenged to ensure that consecutive samples  $(X_n(t_i), X_n(t_i + T))$  and  $(X_n(t_{i+1}), X_n(t_{i+1} + T))$  are independent samples of the stochastic process  $X_n$  with respect to the time separation *between pairs*, which

is  $t_{i+1} - t_i$ .

To address this issue, consider a Gaussian walker on a line with position at time  $t$  distributed as  $P(X_t) \sim \mathcal{N}(0, t\sigma^2) = \frac{1}{\sigma\sqrt{2\pi t}} e^{-\frac{1}{2}\left(\frac{t}{\sigma\sqrt{t}}\right)^2}$  and which starts at  $X_0 = 0$ . We will compute the exact mutual information (nonlinear correlations) between the positions of the Gaussian walker at time  $t_i$  and time  $t_j$  with  $j \geq i$  to illustrate how the mutual information between two realizations of one stochastic process is affected by the time separation between these realizations. Recall that the mutual information between two variables  $X$  and  $Y$  is given by  $MI(X; Y) = \int \int P_{X,Y}(x, y) \cdot \log \left( \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} \right) dx dy$ . We know the marginal densities  $P(X_{t_i}) \sim \mathcal{N}(0, t_i\sigma^2)$  and  $P(X_{t_j}) \sim \mathcal{N}(0, t_j\sigma^2)$  are both Gaussian, and it is known that the mutual information between two Gaussians is  $MI(X, Y) = -\frac{1}{2} \log(1 - \rho^2)$  where  $\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$  [32]. Therefore, we can find the exact mutual information between  $X_{t_i}$  and  $X_{t_j}$  if  $Cov(X_{t_i}, X_{t_j})$  is known. In discrete time, we can compute the Gaussian walker's path using the rule  $X_t = X_{t-1} + \epsilon_t$  where  $\{\epsilon_{t_k}\}$  are a series of independent identically distributed random variables:  $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ . Since  $X_{t-1} = X_{t-2} + \epsilon_{t-1}$ , we find that  $X_t = X_{t-2} + \epsilon_t + \epsilon_{t-1}$ . The recursive relation can be applied  $t - 1$  times to rewrite  $X_t$  as  $X_t = X_0 + \epsilon_t + \epsilon_{t-1} + \dots + \epsilon_2 + \epsilon_1$ . Thus, we can write our realizations as  $X_{t_i} = X_0 + \epsilon_{t_i} + \epsilon_{t_i-1} + \dots + \epsilon_2 + \epsilon_1$  and  $X_{t_j} = X_0 + \epsilon_{t_j} + \epsilon_{t_j-1} + \dots + \epsilon_{t_i} + \dots + \epsilon_2 + \epsilon_1$ . From here, we compute

$$Cov(X_{t_i}, X_{t_j}) = Cov(X_0 + \epsilon_{t_i} + \epsilon_{t_i-1} + \dots + \epsilon_2 + \epsilon_1, X_0 + \epsilon_{t_j} + \epsilon_{t_j-1} + \dots + \epsilon_{t_i} + \dots + \epsilon_2 + \epsilon_1).$$

Since  $X_0 = 0$ ,

$$Cov(X_{t_i}, X_{t_j}) = Cov(\epsilon_{t_i} + \epsilon_{t_i-1} + \dots + \epsilon_2 + \epsilon_1, \epsilon_{t_j} + \epsilon_{t_j-1} + \dots + \epsilon_{t_i} + \dots + \epsilon_2 + \epsilon_1).$$

The covariance is distributive:  $Cov(U + V, W) = Cov(U + W) + Cov(U + V)$ , which we can apply:

$$Cov(X_{t_i}, X_{t_j}) = \sum_{k=0}^{t_i} \sum_{\ell=0}^{t_j} Cov(\epsilon_{t_k}, \epsilon_{t_\ell}).$$

Because  $\{\epsilon_{t_k}\}$  are i.i.d., we know that  $Cov(\epsilon_k, \epsilon_\ell) = 0$  for  $k \neq \ell$ , which reduces our covariance calculation to:

$$Cov(X_{t_i}, X_{t_j}) = \sum_{k=0}^{t_i} Cov(\epsilon_k, \epsilon_k) = \sum_{k=0}^{t_i} Var(\epsilon_k) = \sum_{k=0}^{t_i} \sigma^2 = t_i \sigma^2.$$

Thus,  $\rho = \frac{t_i \sigma^2}{\sigma \sqrt{t_i} \sigma \sqrt{t_j}} = \sqrt{\frac{t_i}{t_j}}$ , and the resulting mutual information is given by:

$$MI(X_{t_i}, X_{t_j}) = -\frac{1}{2} \log \left( 1 - \frac{t_i}{t_j} \right), \quad j \geq i.$$

As expected, the mutual information is infinite in the case where  $t_i = t_j$ , and decreases to zero in the limit as  $t_j \rightarrow \infty$  for fixed  $t_i$ . This relationship is shown in Fig. 2.9.

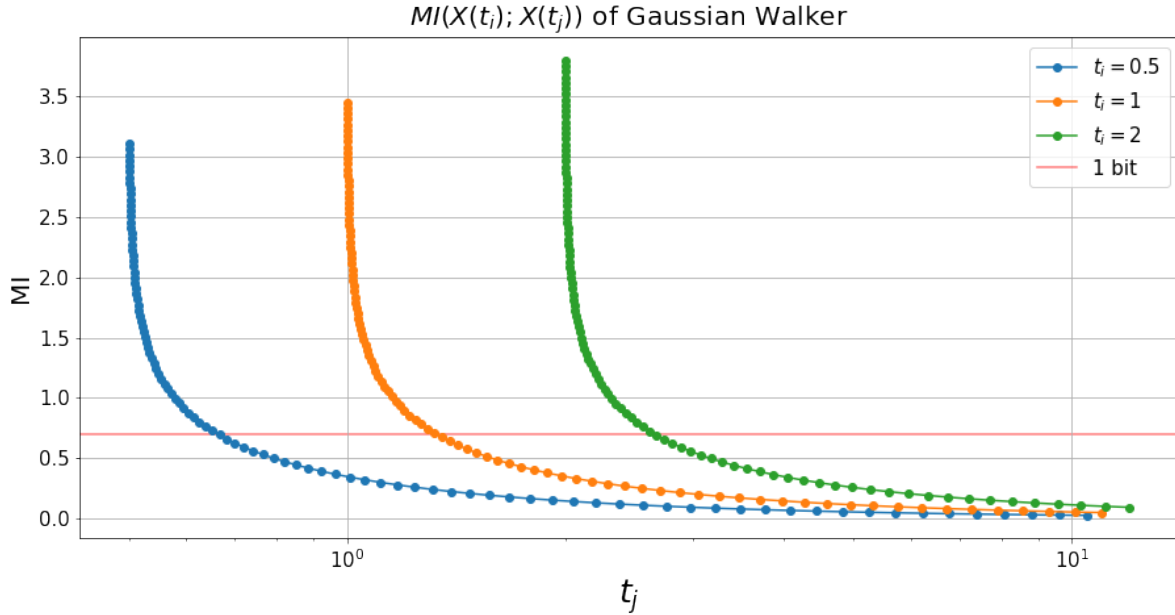


Figure 2.9: For a Gaussian walker, the position at  $t_i$  is correlated to the position at later time  $t_j$ . The mutual information  $MI(X(t_i); X(t_j))$  decreases as the separation between  $t_i$  and  $t_j$  decreases. This decay takes longer for larger  $t_i$  since information is being added to the system at each step.

We find that even for a memoryless Gaussian walker, there is non zero mutual information and therefore nonzero nonlinear correlations between  $X_{t_i}$  and  $X_{t_j}$ . However, for  $t_j$  sufficiently larger than  $t_i$ , these correlations decay exponentially to zero. We introduce a parameter  $W$  which we will enforce as the average window of separation between consecutive sampled times  $t_i$  and  $t_{i+1}$ . Below, we sketch the algorithm for this sampling procedure, in which the set of time separations  $\{t_{i+1} - t_i\} = \{W_i\}$  is a random variable. Realizations of this random variable,  $\{W_i\}$ , are iid with mean  $W$ , the true window size. We illustrate this fact in Fig. 2.11.

We illustrate this sampling scheme in Fig. 2.12, where we take the path data of a single path  $X_n$  from the ensemble (shown in Fig. 2.7). This path is then sampled at discrete times  $\{t_i\}$  with average separation  $W$  and

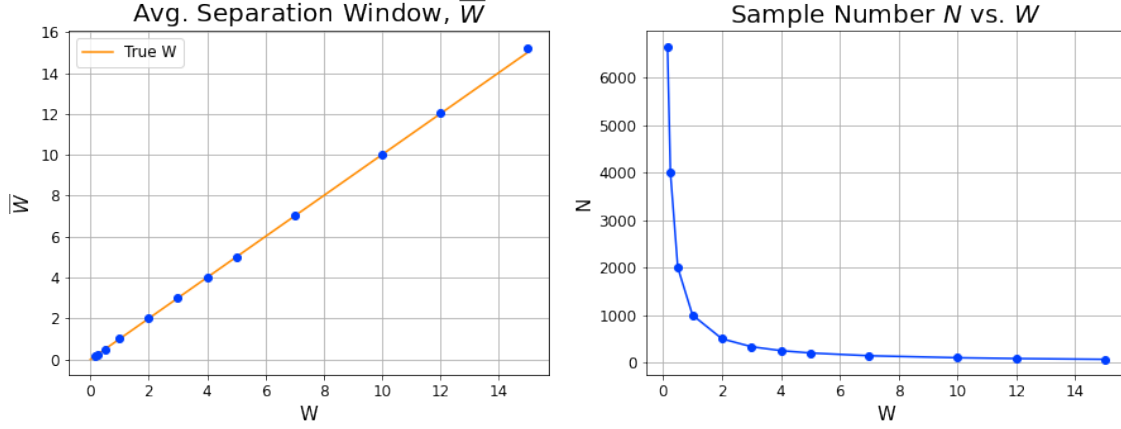


Figure 2.10: Using test data of length 100,000 we compute the average separation between samples  $\bar{W}$  with standard errors. (Each data point is the average of 10 repetitions, with standard errors that are smaller than the marker size.) The true window  $W$  is shown in orange. In the second panel, we illustrate the sample size  $N$  as a function of  $W$ . Although more samples usually yield more accurate estimates, the KNN mutual information estimator has been shown to work well for small sample sizes. In [32], the smallest sample sizes considered are  $O(10^2)$ . Despite lower sample sizes in our case, we find our results to be consistent, although with higher standard errors.)

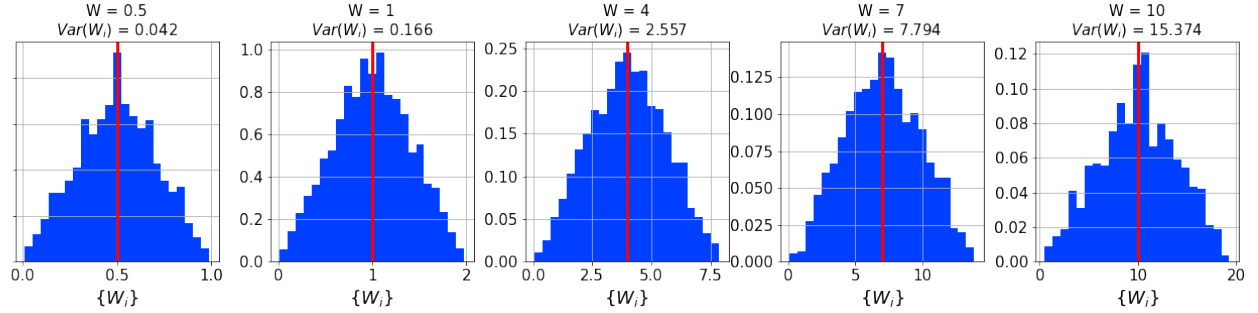


Figure 2.11: As  $W$  increases, the variance of the distribution of  $\{W_i\}$  increases, although the mean value (true value  $W$ ) remains the same.

again at delayed times  $\{t_i + T\}$  to generate the random sample:  $\{(X_n(t_0), X_n(t_0 + T)), X_n(t_1), X_n(t_1 + T)), \dots, X_n(t_f), X_n(t_f + T)), \}$ .

Proper choice of  $W$  will ensure that we sample on a timescale that removes (or suppresses) these dynamical correlations. If  $W$  is too small, then the reported MI will include the effect of these dynamical correlations and the reported mutual information might be artificially inflated. Since we expect that past and future correlations decay as the time between them increases, overly large  $W$  will merely reflect that samples that are very far apart in time do not influence one another. Thus, we want  $W$  to be large enough to overlook dynamical correlations, but small enough to capture a signal. To justify the choice of  $W$ , we might naturally

choose  $W$  to be the timescale on which the (linear) autocorrelation of  $X_n$  decays to zero. We will assess how we might find sufficiently large  $W$  in a later section. We note here that  $T$  need not be constrained by  $W$ , as demonstrated in the Fig. 2.12.

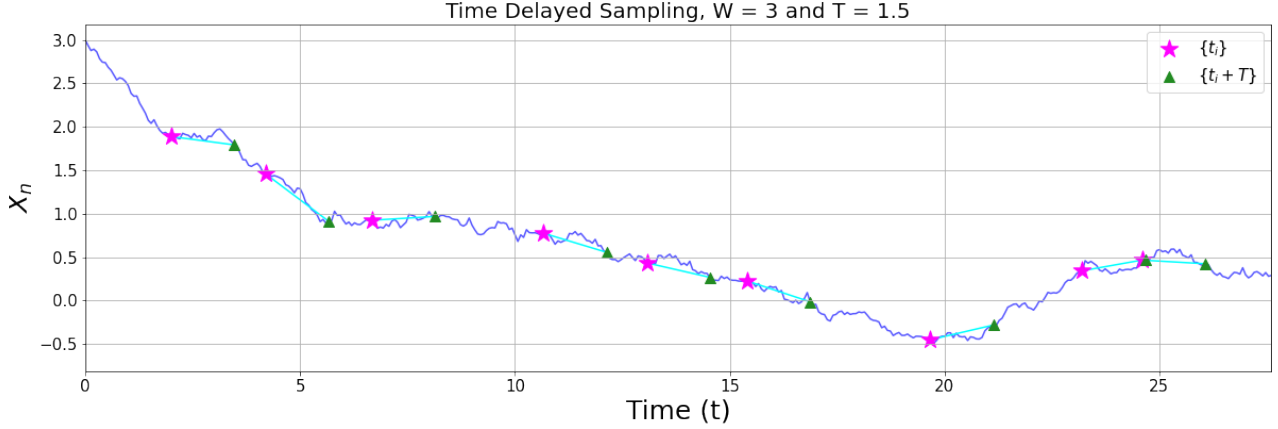


Figure 2.12: Using a single Ornstein-Uhlenbeck trajectory  $X_n(t)$  from  $t = 0$  to  $t = 27$ , we select magenta initial samples with times  $\{t_i\}$ , which must be separated by windows  $W_i$  satisfying  $\langle W_i \rangle \sim W$ . The time delay  $T = 1.5$  is then used to select the time-delayed curvatures (green samples).

For the methods described above, we have focused on the time delayed self mutual information of a single agent,  $X_n$ . The self mutual information framework will be used to investigate single path statistics of the self-avoidant swimming droplets. To study the pairs of golden shiners, we can apply these same methods to the time delayed mutual information *between* two agents,  $X_n$  and  $X_m$ , which is  $MI(X_n(t); X_m(t + T))$ . (In our case, we will only have two agents, so we can write this as  $MI(X(t), Y(t + T))$ ). We can switch which agent we sample on a time delay allowing us to study the differences between  $MI(X_n(t); X_m(t + T))$  and  $MI(X_n(t + T); X_m(t))$ . Differences between these two values will allow us to use mutual information to suggest directional information flow *without* conditioning on the past history of any variable, as is required to compute transfer entropy or causation entropy.

### 2.3.1 Limitations and Advantages

Throughout our discussion of estimating the time delayed self mutual information ( $MI(X_n(t); X_n(t + T))$ ) and time delayed mutual information between agents ( $MI(X_n(t); X_m(t + T))$ ), a fundamental assumption is that the stochastic process described by  $X_n$  is sampling from a stationary distribution. In the case of the Gaussian walker (whose position is not taken from a stationary distribution), more information is

added to the system with each time step since each new position is formally dependent on all past positions. This is reflected in Fig. 2.9, where the mutual information curves increase with increasing  $t$ . Therefore, for sufficiently large  $t$  any fixed window size  $W$  would eventually become too small for the dynamical correlations to decay. Thus we find that although the Gaussian walker nicely illustrates the effect of separating consecutive samples by average window size  $W$ , our methodology would not work well for estimating the time delayed self mutual information of the entire Gaussian walker path.

In the case of our unconfined swimming droplets, the positions will not sample the configuration space well at long times. To work around this, in a later section we will introduce a new random variable that will sample a bounded configuration space well at long times. In the case of the golden shiner pairs, both the confined experimental domain (an annular tank) and the long experimental trajectories will contribute to the data sampling the positional configuration space (all positions within the tank) well at long times.

A key advantage of the data we use in this work is its length. Our model generated swimming droplet trajectories are substantially longer than tracked trajectories of experimental swimming droplets, and the experimentally collected golden shiner path data is substantially longer than animal path data collected in the wild. Additionally, unlike data collected in the wild, we allow our experimental golden shiners to equilibrate in their environment before collecting data. Our method of using time series data to calculate time delayed mutual information is also underexamined in the literature. As mentioned previously, transfer entropy and causation entropy have been used to infer leader-follower dynamics from path data of animal pairs, but such methods have significant drawbacks both computationally and in terms of the certainty [10], [58], [47], [45].

## CHAPTER 3

### Exploration of a Model for Swimming Droplets

#### 3.1 Introduction

A hallmark feature of microscale active systems is a ballistic movement, or “swimming” that when interrupted by random and frequent directional changes gives rise to enhanced diffusion [11, 49, 53, 31]. The biological advantage of enhanced diffusion is greater exploration of an area in a shorter period of time when compared to passive diffusion. Consequently, biological effective diffusion and other single-particle emergent behaviors such as micro-scale transport, bacterial motion, and cell migration patterns and their biomimetic applications are research areas of great interest [80].

In parallel, complete understanding of these phenomena via mathematical modeling provides design inspiration and permits cost-effective testing of novel systems; the most common model for active particles is the active Brownian particle (ABP) model. ABP combines directed motion resulting from a velocity dependent on the amount of available energy or “fuel” with a rotational diffusion dependent on a defined persistence timescale, resulting in enhanced diffusion at time scales longer than the correlation time of the rotational diffusion [31]. This model of competing ballistic and diffusive motion accurately predicts the enhanced diffusion of many experimental systems, such as those found in Refs. [36, 72, 34, 11, 64, 18, 31].

We are interested in the additional effect of spatio-temporal memory observed in slowly-dissolving autophoretic droplets [44, 35, 33, 30], in addition to the persistence memory seen in ABP. As these autophoretic droplets interact with the surfactant suspension, the particular physics induces a self-avoidant memory response. Above a critical surfactant concentration, the leaking oily solute from the droplets is taken up into empty micelles. This creates local heterogeneities in the surfactant concentration, which induce Marangoni stresses that cause the droplets to spontaneously swim in the direction of highest surfactant concentration. This process continues as the droplets move, leaving behind a diffusing wake of solute-filled micelles and thereby a trail of depleted surfactant concentration. It is precisely the fact that the diffusion of the micelles and surfactant is slow relative to the velocity of the droplets that causes self-avoidant motion as the

droplets encounter gradients of solute concentration at the droplets' past locations that have not yet diffused away. These past-history gradients induce Marangoni stresses that cause the droplets to move towards higher surfactant concentrations and therefore away from their past locations.

Despite being too large for the effects of thermal noise to be visible, the ballistic motion of the autophoretic experimental droplets is still punctuated by randomized directional changes, producing random-walk-like behavior. Such changes in direction reflect a transition between a dipolar (swimming) and a quadrupolar (stopped) hydrodynamic mode and the average frequency of these re-orientation events increases with Péclet number, droplet size, and the viscosity of the surrounding suspension [30]. While this run-and-tumble-like behavior produces an enhanced diffusion that is consistent with the ABP model for the experimental parameters considered in [33], we seek an understanding of the additional self-avoidant memory effect at play, particularly on the enhanced diffusion.

Motivated by the experimental system, we employ a model with a tunable memory response (which we distinguish from directional persistence) that qualitatively captures the essential features of the droplets and ignores the details of Marangoni stresses and hydrodynamic effects. In this model, the particle is a mobile source of diffusing surfactant that descends its self-produced concentration gradient, resulting in a sustained “swimming” state and self-avoidant memory tied to the diffusion timescale. To reproduce the coarse-grained effect of the random reorientations after each switch from the quadrupolar hydrodynamic mode of the experimental particles, we introduce thermal-like noise into the droplet’s equation of motion. This results in enhanced diffusion that intuitively one might expect the encoded self-avoidant memory to amplify as the particle evades its own past locations. However, we find the opposite: a suppression of enhanced diffusion over that predicted by an ABP with the same velocity and orientational persistence. We find evidence of *transient self-caging* as a possible explanation for this behavior.

In this chapter, we begin in Sec. 3.2 by presenting the mathematical details of this model for self-avoidant swimming droplets. We investigate the memory effects of these model swimmers at long time scales by comparing the mean square displacement (MSD) to that of ABPs with the same velocity and orientational persistence in Sec. 3.3. To make these comparisons, we analytically derived an expression for the velocity in our model and numerically compute its orientational persistence timescale. We find that the equivalent ABP *overestimates* the enhanced diffusion of the model self-avoidant droplets, which we attribute to an unexpected side-effect of self-avoidant memory: transient self-caging. In Sec. 3.4 we further investigate the parameter space of the model, finding that with fixed noise strength, there is a limited regime of self-avoidant-memory

strength within which enhanced diffusion is observable; the zero-memory limit of our model is not ABP. We conclude the paper in Sec. 3.5.

### 3.2 A Model for Self-avoidant Memory

Motivated by the experimental system described previously, we propose a coupled model of a diffusion partial differential equation (PDE) for the surfactant concentration  $c(\mathbf{x}, t)$  and a stochastic differential equation (SDE) for the particle's location  $\mathbf{X}(t)$ . These equations are

$$\partial_t c(\mathbf{x}, t) = D\Delta c(\mathbf{x}, t) + \alpha DR^2 \delta_R(\mathbf{x} - \mathbf{X}(t)), \quad (3.1a)$$

for  $\mathbf{x} \in \Omega \subseteq \mathbb{R}^2$ ,  $t \geq 0$ , and

$$d\mathbf{X}(t) = -\beta R \left( \int_{\Omega} \delta_R(\mathbf{x} - \mathbf{X}(t)) \nabla_{\mathbf{x}} c(\mathbf{x}, t) d\mathbf{x} \right) dt + \sqrt{\sigma} d\mathbf{W}(t), \quad (3.1b)$$

with prescribed initial conditions ( $c(\mathbf{x}, 0) = 0$  and  $\mathbf{X}(t) = \mathbf{0}$  unless otherwise noted) and boundary conditions (reflecting boundary conditions on  $\partial\Omega$  unless otherwise noted).

Equation (3.1a) is a diffusion equation with diffusivity  $D$  and a source term at the particle's current location. The time evolution of the concentration field holds the temporally-decaying memory of the particle's spatial history. We note that the inclusion of diffusion on the source term differs from similar models for chemoattractive forces [62, 25, 38]. Motivation for this decision and the resulting effects are discussed later in this section.

In the source term with rate  $\alpha D$ , we introduce a “size”  $R$  to the particle using the radially-symmetric mollified delta function,  $\delta_R(\mathbf{x} - \mathbf{X}(t)) = \frac{1}{2\pi R^2} e^{-\frac{|\mathbf{x} - \mathbf{X}(t)|^2}{2R^2}}$ . (For the treatment of the particle as a point source with the Dirac delta function, see App. B.2; interestingly the particle does not swim.) As the droplet releases oily solute from its membrane located at  $|\mathbf{x} - \mathbf{X}(t)| = R$ , this Gaussian emission pattern with standard deviation  $R$  approximates the physical boundary of the particle while being more numerically and analytically tractable and does not require imposing a moving boundary condition on the concentration field to exclude the particle's interior. Since the particle's boundary does not physically exist in this model, Eq. (3.1a) also

ignores the subtle effects the induced advection along the particle’s surface has on the concentration gradient, as detailed in Ref. [44]. We also keep  $R$  fixed in time, thus we ignore depletion effects.

Equation (3.1b) is a modified Langevin equation for a Brownian particle in a force field in the strong friction limit. We again mimic the size of the particle by convolving the gradient of the concentration field with the mollified delta function. We define the particle’s response strength to the concentration gradient to be  $\beta R$ .

As is conventional for an overdamped Langevin equation,  $\mathbf{W}$  is a two-dimensional Wiener process scaled by the noise strength  $\sqrt{\sigma}$ . Recall, the experimental particles are athermal; this noise is to reproduce the stochasticity introduced by local fluctuations in the surfactant gradient that lead to re-orientations of the experimental droplet’s swimming direction after each switch from a quadrupolar to bipolar mode. As the frequency of these re-orientation events depends on Péclet number, droplet size, and the viscosity of the surrounding suspension [30], the parameter  $\sigma$  would likely be linked to other model parameters like diffusion  $D$ , droplet “size”  $R$ , and response to the concentration gradient  $\beta$ . As we wish to keep noise effects constant to isolate the effects of self-avoidant memory in the present study, we ignore these possible dependencies in this study.

The stated model in Eq. (3.1) articulates the explicit relationship between the evolving concentration field and the particle trajectories. As the particle moves, its emissions induce changes in the local concentration field and it leaves behind diffusing physical evidence of its trajectory. Thus, the historical information or memory of the particle’s past locations is contained within the current state of the evolving concentration field. The memory encoded in the concentration field allows each particle to “remember” where it has been (hotspots in the concentration field) and avoid its past trajectory with response strength decreasing as the time lag increases. Via integration of the whole gradient field at each time point, the particle becomes spatially omniscient as it moves with dependence on the affecting forces from every spatial point on the domain. In time, the particles are pseudo-omniscient as their ability to “see” into the past through interaction with the concentration gradient diminishes exponentially in time. This unique behavior abolishes time-reversal symmetry although the coupled configuration in Eq.(3.1) is Markovian since there is no explicit dependence on the trajectory’s past steps.

To limit the number of parameters under investigation, we nondimensionalize Eq. (3.1). We choose  $R$  as a natural length scale and non-dimensionalize  $c$  without any scaling for simplicity. Temporarily, we leave

time scale  $T$  arbitrary. Under the scalings  $\mathbf{y} = \frac{\mathbf{x}}{R}$ ,  $\mathbf{Y} = \frac{\mathbf{X}}{R}$ ,  $\tau = \frac{t}{T}$ , and  $\mathbf{B} = \frac{\mathbf{W}}{\sqrt{T}}$ , we arrive at

$$\partial_t c(\mathbf{y}, t) = \mu \Delta c(\mathbf{y}, t) + \mu \phi \exp \left[ -\frac{|\mathbf{y} - \mathbf{Y}(t)|^2}{2} \right], \quad (3.2a)$$

for  $\mathbf{y} \in \Omega$ ,  $t \geq 0$ , and

$$d\mathbf{Y}(t) = -\nu \left( \int_{\Omega} \exp \left[ -\frac{|\mathbf{y} - \mathbf{Y}(t)|^2}{2} \right] \nabla_{\mathbf{y}} c(\mathbf{y}, t) d\mathbf{y} \right) dt + \sqrt{\epsilon} d\mathbf{B}(t) \quad (3.2b)$$

where  $c$ ,  $t$  and  $\Omega$  are re-used for their non-dimensional versions for convenience. We have mapped the dimensional parameters as follows:  $D \rightarrow \mu = \frac{DT}{R^2}$ ,  $\alpha \rightarrow \phi = \frac{\alpha R^2}{2\pi}$ ,  $\beta \rightarrow \nu = \frac{\beta T}{2\pi R}$ , and  $\sigma \rightarrow \epsilon = \frac{\sigma T}{R^2}$ .

We note that a typical time scale for the diffusion equation is  $T = \frac{L^2}{D}$ . Although traditional, this choice would prevent us from seeing directly the effects of changing  $D$ , which encodes the memory timescale. Increasing diffusivity would contract time such that the past-history wake of the particle would adjust to decay at the same rate. Thus, to observe the effects of this memory-encoding diffusivity, we choose to fix the stochastic diffusivity  $\epsilon = 0.75^2$ , thereby choosing  $T = 0.75^2 \frac{R^2}{\sigma}$ . Keeping the value of  $\sqrt{\epsilon}$  fixed at  $\frac{3}{4}$  was a convenient choice made to maintain balance between the stochastic effects, controlled by  $\sqrt{\epsilon}$ , and the deterministic effects of swimming as well as self-avoidant memory, controlled by  $\nu$  and  $\mu$ .

We can simplify the system by taking the Fourier transform and solving the PDE (3.2a) on an infinite domain,  $\Omega = \mathbb{R}^2$ , explicitly. Incorporating this solution into the SDE (3.2b) we arrive at the mathematically equivalent system for the particle in an infinite domain

$$d\mathbf{Y} = \frac{\pi}{2} \mu \nu \phi \int_0^t \exp \left[ -\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(1 + \mu(t-s))} \right] \cdot (1 + \mu(t-s))^{-2} (\mathbf{Y}(t) - \mathbf{Y}(s)) ds dt + \sqrt{\epsilon} d\mathbf{B}; \quad (3.3)$$

see App. B.1 for details. This non-Markovian SDE explicitly reveals the dependence on all the particle's previous locations via integration in time; the exponential kernel decays in both time and space, revealing  $\mu^{-1}$  to be the self-avoidant memory timescale. In this way the model contains a self-avoidant memory, one of the key features of the experimental droplets, with controllable timescale  $\mu^{-1}$ . In addition to producing a self-sustained swimming state, the form of this force also allows the droplets to hover above a bottom plate with the addition of gravity to the model (see App. B.3 for details) in much the same way that the experimental droplets do [44].

The formulation of the model in Eq. (3.3) is convenient for simulation since it does not require solving the PDE on a large domain to capture long-time dynamics. It additionally removes the integral in the SDE over  $\mathbb{R}^2$  and replaced it with an integral over  $t$ . We integrate Eq. (3.3) in time with the Euler-Maruyama method while using Simpson's rule to integrate the memory kernel at each step. This algorithm is a first order method.

The limiting behavior of these two equivalent systems, Eq. (3.2) and (3.3), foreshadows their distinction from the active Brownian model since it reveals that removing the distinguishing feature of memory by taking  $D \rightarrow \infty$  will not reduce our model to ABP. The parameter  $D$  was added to the source term in Eq. (3.1a) to achieve balance between the rate at which the oil diffuses and the rate at which the oil is expelled in this limit. (If instead the source term remained constant relative to  $D$ , then it would effectively vanish in the limit of  $D \rightarrow \infty$ .) The nondimensional parameter  $\mu$  therefore appears on the source term in Eq. (3.2a), and  $\partial_t c(\mathbf{y}, t) \rightarrow \infty$  as  $\mu \rightarrow \infty$ . To leading order, the concentration field satisfies the Poisson equation

$$\Delta c(\mathbf{y}) = -\phi \exp \left[ -\frac{|\mathbf{y} - \mathbf{Y}(t)|^2}{2} \right] \quad \mu \rightarrow \infty. \quad (3.4)$$

The concentration field is now memory-less since it instantaneously equilibrates as the particle moves. On an infinite domain, the solution to Eq. (3.4) will be radially symmetric around the particle's location, and therefore the integral in Eq. (3.2b) will always be zero. As a result, particles experience motility solely from thermal fluctuations, namely simple Brownian motion.

Also noteworthy is the “full memory” limit of Eq. (3.2a) which is  $\partial_t c \rightarrow 0$  as  $\mu \rightarrow 0$ . The concentration field remains fixed at its initial conditions as the source term and the diffusion term vanish in this limit. The particle experiences thermal fluctuations while responding to the concentration gradient of the fixed concentration landscape, thereby statistically preferring concentration minima. The steady state (if one exists) would be almost solely determined by the initial topography of  $c(\mathbf{y})$  and the relative size of  $\epsilon$ . In fact, the entire coupled system in Eq. (3.2) reduces to simple memoryless Brownian motion

$$d\mathbf{Y} = \sqrt{\epsilon} d\mathbf{B} \quad \mu \rightarrow 0 \quad (3.5)$$

in this limit, as the source term which encodes the memory vanishes. This is consistent with Eq. (3.3) which also reduces to simple Brownian motion in the same limit  $\mu \rightarrow 0$  when it is assumed that the initial

concentration field is constant. Therefore we focus our study on intermediate range of  $\mu$  where the effects of the noise, the swimming, and the memory are all observable. The limiting behaviors of our model as described above can all be traced back to the addition of a second  $\mu$  on the source term. As stated previously, inclusion of a diffusive scaling on the source term was required to ensure that the source term remained in the limit as  $D \rightarrow \infty$ , in the hope that the model would revert to active Brownian motion with no self-avoidant memory. In our results, we discuss the role of this diffusive scaling in generating previously unseen behaviors in this class of models.

### 3.3 Comparative Analysis

Numerically simulated trajectories of the coupled model given by Eq. (3.2) are shown in Fig. 3.1. These trajectories illustrate the main features of active matter: a swimming velocity with a slowly diffusing direction. Increasing  $\nu$ , and therefore the response to the concentration gradient causes the particle to swim faster, shown in Fig. 3.1a, while increasing  $\mu$ , and therefore shortening the timescale of the diffusion (decreasing the memory), has a secondary effect on the velocity, but also causes the particles to turn faster, shown in Fig. 3.1b. An increase in turning frequency was also observed experimentally in [33] as surfactant concentration was increased, prompting a transition from ballistic motion to diffusion. (See Figure 2 in [33]. Recall the Marangoni effect which causes the droplets to “search” for areas of higher surfactant concentrations, while the droplets simultaneously modify the local concentration. )

We seek to look beyond the combined effects of swimming and random directional changes in producing enhanced diffusion and understand the additional effects of self-avoidant memory. Specifically, we compare our model to ABP given by the equations

$$dX = V \cos(\theta(t))dt + \sqrt{\epsilon}dW_x \quad (3.6a)$$

$$dY = V \sin(\theta(t))dt + \sqrt{\epsilon}dW_y \quad (3.6b)$$

$$d\theta = \frac{1}{\sqrt{\tau}}dW_\theta \quad (3.6c)$$

where  $\sqrt{\epsilon}$  is the strength of the additive noise in each spatial component (consistent with the model in Eq. (3.3)),  $V$  is the swimming velocity, and  $\tau$  is the persistence timescale of the rotational diffusion [33, 31, 42].

These latter two parameters do not explicitly appear in our model; we will compute them and compare the MSD of the two models to understand the effects of self-avoidant memory on enhanced diffusion.

In Sec. 3.3.1 we present an analytic equation that is numerically solved for the velocity of the swimming solution to Eq. (3.3). This velocity is consistent with the intermediate ballistic regime of the MSD, computed numerically for Eq. (3.3) and given by

$$\mathbb{E}[\mathbf{X}(t)^2] = 4V^2\tau^2 \left[ 2 \left( e^{-\frac{t}{2\tau}} - 1 \right) + \frac{t}{\tau} \right] + 2\epsilon t \quad (3.7)$$

for Eq. (3.6). In Sec. 3.3.2 we determine  $\tau$  by numerically computing the orientation correlation function but find that the memory induced from modifying the environment causes a reduced effective diffusion as compared to ABP with identical angular persistence.

### 3.3.1 Intermediate Time Scales: Ballistic Motion

Unlike active Brownian models, the proposed model has a non-explicit intrinsic velocity; directed motion at this velocity may become observable at intermediate time scales under appropriate conditions for  $\nu$  and  $\mu$ . To find an analytic form for the velocity, we seek a deterministic constant velocity (“steady state”) solution to the combined model Eq. (3.3). Without loss of generality, suppose  $\mathbf{Y}(t) = \langle Vt, 0 \rangle$ ; thus  $V$  must solve

$$\frac{d\mathbf{Y}}{dt} = V = \frac{\pi}{2} \mu \nu \phi \int_{-\infty}^t \exp \left[ -\frac{|Vt - Vs|^2}{4(1 + \mu(t - s))} \right] \cdot (1 + \mu(t - s))^{-2} (Vt - Vs) ds. \quad (3.8)$$

Under the transformation  $z = \mu(t - s)$ , the constant velocity  $V$  therefore satisfies

$$1 = \frac{\pi}{2} \frac{\nu}{\mu} \phi \int_0^\infty \frac{z}{(1 + z)^2} \exp \left[ -\left( \frac{V}{\mu} \right)^2 \frac{z^2}{4(1 + z)} \right] dz. \quad (3.9)$$

For each value of  $\frac{\nu}{\mu}$ , we solve for the value of  $\frac{V}{\mu}$  that makes the above integral equal to 1 numerically in Python with `scipy`. Under the change of variables  $x = \frac{2}{\pi} \arctan(z)$ , we map the domain  $(0, \infty)$  to  $(0, 1)$  for ease of numerical integration. The resulting monotonically increasing dependence of  $\frac{V}{\mu}$  on  $\frac{\nu}{\mu}$  is plotted as the solid black line in Fig. 3.2b. Alternately, we can select  $\mu$  and  $V$  and compute  $\nu$  to satisfy Eq. (3.9).

We can directly compare this theory to ABP on the timescale at which ballistic motion is dominant. It is evident from Fig. 3.2a that the ballistic portion of the simulated MSD aligns with the computed velocity from

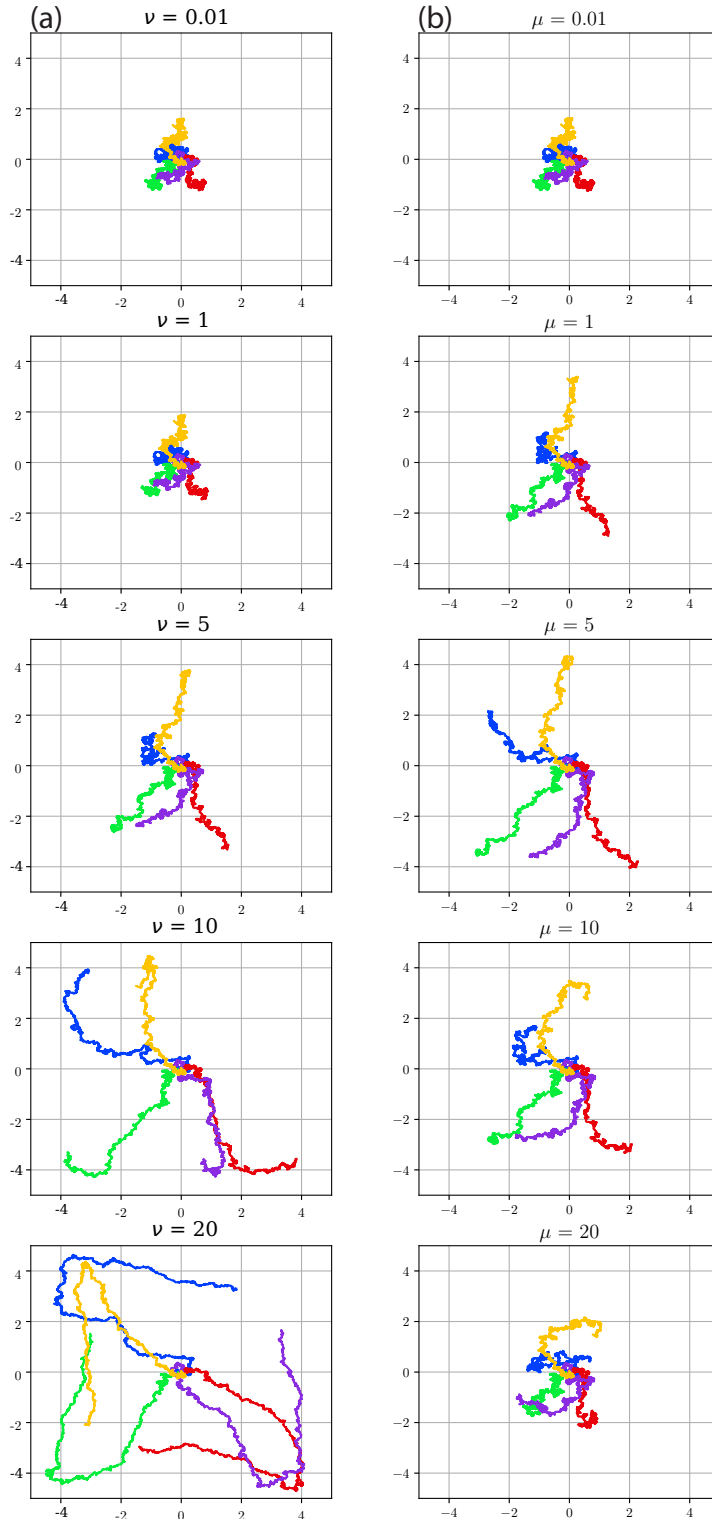


Figure 3.1: Numerical simulations of Eq. (3.2) were carried out using a forward-time centered-space finite difference scheme for the PDE and the Euler-Maruyama method for integrating the SDE where the trapezoid rule was used to compute the integral therein. (Caption continued on next page.)

Figure 3.1: We confine  $\mathbf{X}(t)$  to a box  $B = \{x \in (-5, 5), y \in (-5, 5)\}$  with insulating boundary conditions such that  $\frac{\partial c(\mathbf{x}, t)}{\partial t} = 0 \quad \forall \mathbf{x} \in \{\partial B\}$ . This has the effect of reflecting the particle back into the domain when it reaches the boundary. The initial condition is  $c(\mathbf{x}, 0) = 0$ . Note that these trajectories are visually indistinguishable from active Brownian motion. (a) For  $\phi = 1$ ,  $\mu = 5$  and a noise level of  $\sqrt{\epsilon} = 0.75$ , we see the dominant effect of  $\nu$  which is to increase the velocity. (b) For  $\phi = 1$ ,  $\nu = 7$ , and a noise level of  $\sqrt{\epsilon} = 0.75$ , we see the effects of  $\mu$  which primarily increases the turning frequency and has a secondary effect on velocity.

Eq. (3.9). At such small times, the MSD of ABP asymptotically reduces to

$$\mathbb{E}[\mathbf{X}(t)^2] \sim V^2 t^2 + 2\epsilon t \quad (3.10)$$

as  $t \rightarrow 0$  (see App. B.4 for details.) Fitting  $V$  from the ballistic portion of the MSD of our particles is also in good agreement with the theory, as shown in Fig. 3.2b.

We point out that the existence of an observable ballistic regime in the MSDs from our model requires a sufficient swimming velocity  $V$  to dominate the additive noise. In the ABP model, this can be guaranteed by changing the stated parameter  $V$ , whereas in our model, there must be consideration for the parameters  $\mu$  and  $\nu$  due to the explicit functional relationship  $V = f(\mu, \nu)$  given by Eq. (3.9). To see this functional relationship more clearly, the contours of constant velocity are plotted in Fig. 3.3a and the contours of constant  $\nu$  are plotted in Fig. 3.3c. These figures agree with the limits from Sec. 3.2 in that  $V = f(\mu, \nu) \rightarrow 0$  when taking either  $\mu \rightarrow 0$  or  $\mu \rightarrow \infty$  with fixed  $\nu$ , and the model system Eq. (3.2) or Eq. (3.3) reduces the particle motion to simple Brownian motion. Furthermore, taking  $V \rightarrow 0$  in Eq. (3.8) results in a divergent integral; for the integral to converge, either  $\mu$  or  $\nu$  in the prefactor must also go to zero. The result is no transition to swimming at a small finite value of these parameters. Similarly, the integral also diverges as  $\mu \rightarrow \infty$ . Figure 3.3c most clearly shows the relevant intermediate values of  $\mu$  for which a significant velocity exists and ABP-like motion with a ballistic regime is expected for the model system.

Figure 3.3a most clearly shows that for any given velocity, there exists a minimal  $(\mu, \nu)$  pair. If we interpret this in the context of the coupled model given by Eq. (3.2), it suggests the existence of an optimal response strength and diffusivity pairing which act on the particle to produce directed motion at a specific speed. Moving off of this minimum illustrates the parameter couplings which must balance to keep the particle moving at a given speed. For example, decreasing memory (increasing  $\mu$ ) allows the particle's trail to diffuse faster which weakens local gradients, and thus requires that the response strength to the weakened gradient be increased (increasing  $\nu$ ).

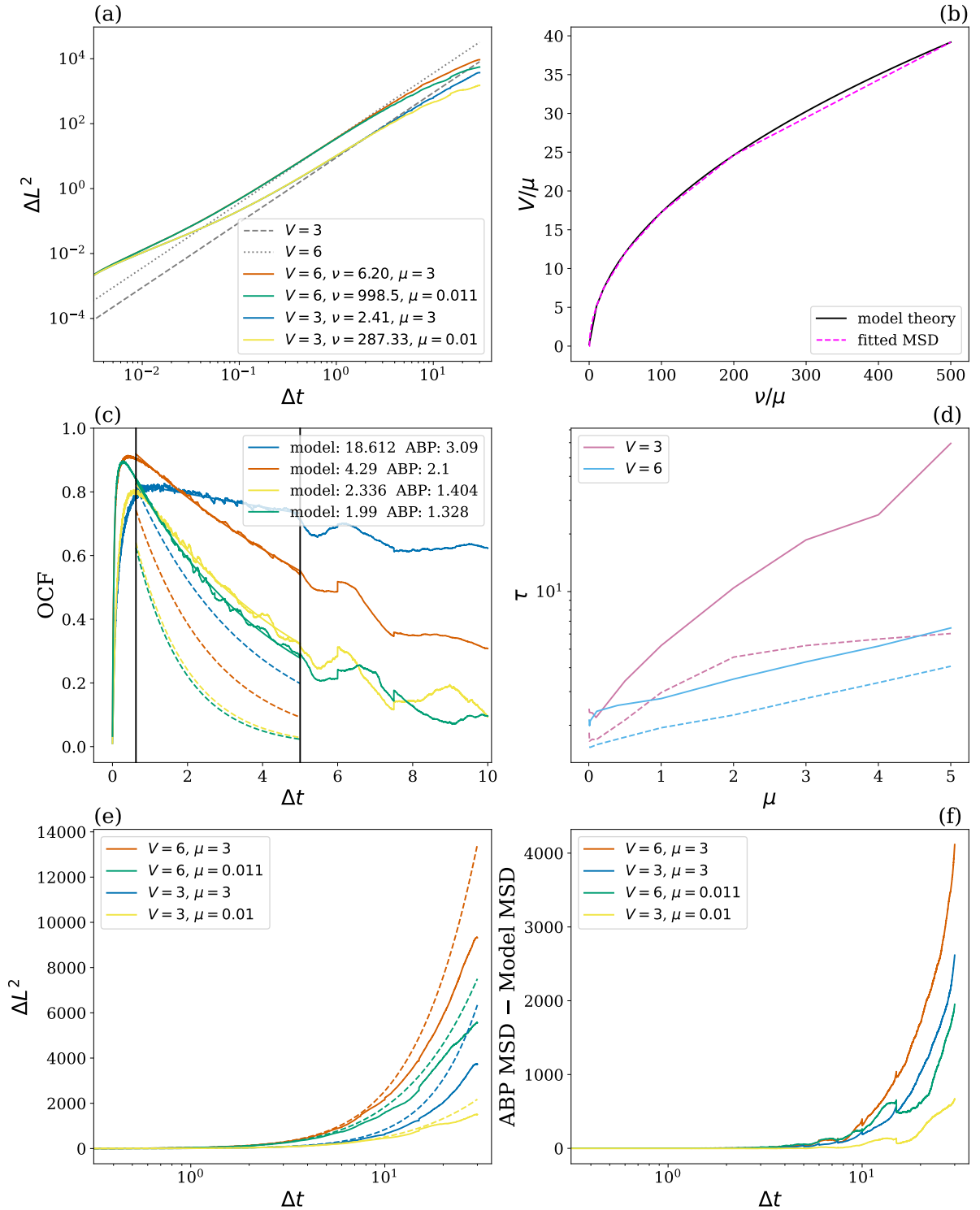


Figure 3.2: (a) Empirically computed MSDs of the model in comparison to benchmark pure ballistic motion. Three distinct regions of the MSD are visible: classical diffusion, directed motion in alignment with the benchmark ballistic lines, and enhanced diffusion where the MSD pulls away from the ballistic motion. Dashed lines correspond to benchmark pure ballistic motion at velocities indicated in legend. Curves in legend are ordered in their appearance in the figure, from top to bottom. (Caption continued on next page.)

Figure 3.2: (b) Collapsed model velocity theory from solving Eq. (3.9) in comparison to fitted values extracted from the ballistic regime of the MSD of the form  $V^2 t^2$ . (c) Model OCFs (noisy solid lines) with fitted  $\tau$  (solid smooth curves) with comparison to  $\tau$  fitted from the enhanced diffusion regime of the MSD consistent with ABP Eq. (3.11) (dashed smooth curves). In the legend are shown values of  $\tau$  for each experiment corresponding to parameters listed in the legend of (a). Legend is ordered as curves appear from top to bottom in the figure. (d) Both fitted  $\tau$  values from full ABP MSD (dashed) and OCF of our model (solid) as they vary over  $\mu$ . (e) Model MSD (solid) and ABP MSD Eq. 3.7 (dashed) under the same  $\tau$  fitted to the model OCF and theoretical  $V$ . Legend ordered by curves from top to bottom. (f) Signed difference between model and ABP MSDs in panel (e). Legend ordered by curves from top to bottom.

### 3.3.2 Long Time Scales: Enhanced Diffusion

Figure 3.2a shows a departure of the MSD from ballistic motion at longer time scales. For ABP, this departure happens at time scales  $t \gg \tau$  for which the MSD (3.7) is asymptotic to

$$\mathbb{E}[\mathbf{X}(t)^2] \sim (4V^2\tau + 2\epsilon)t \quad (3.11)$$

as  $t \rightarrow \infty$ . Particle reorientations that decorrelate with timescale  $\tau$  enhance the diffusion term  $2\epsilon t$  with the term  $4V^2\tau$ .

To estimate  $\tau$  from trajectories given by our model we first numerically compute the normalized orientation correlation function (OCF) which measures the relative angle between consecutive movements. It is given by

$$C(\Delta t) = \left\langle \frac{\mathbf{v}(t) \cdot \mathbf{v}(t + \Delta t)}{|\mathbf{v}(t)| |\mathbf{v}(t + \Delta t)|} \right\rangle_t \quad (3.12)$$

where  $\mathbf{v}(t) = \mathbf{Y}(t) - \mathbf{Y}(t - \Delta t)$  is the directional displacement between times  $t$  and  $t - \Delta t$  (see App. B.5 for details). This function computed for the trajectories is shown in Fig. 3.2c as the noisy solid lines. Note that  $C(\Delta t) \rightarrow 0$  as  $\Delta t \rightarrow 0$  because the motion at such small timescales is dominated by the uncorrelated additive noise. As  $\Delta t$  increases ballistic motion starts to dominate which is reflected in the OCF that approaches values near unity. The portion of  $C(\Delta t)$  displaying exponential decay, due to the transition to enhanced diffusion at even longer  $\Delta t$ , is fit by a single exponential given by  $C(\Delta t) = e^{-\frac{\Delta t}{2\tau}}$  as is consistent with ABP [33, 42]. These fits are shown by the smooth solid lines in Fig. 3.2c and the resulting values of  $\tau$  as a function of  $\mu$  in Fig. 3.2d.

While the model OCF is well fit by an exponential decay, the long time asymptotics of the MSD given by Eq. 3.11 in Fig. 3.2e (dashed lines) shows that ABP substantially overestimates the enhanced diffusion of

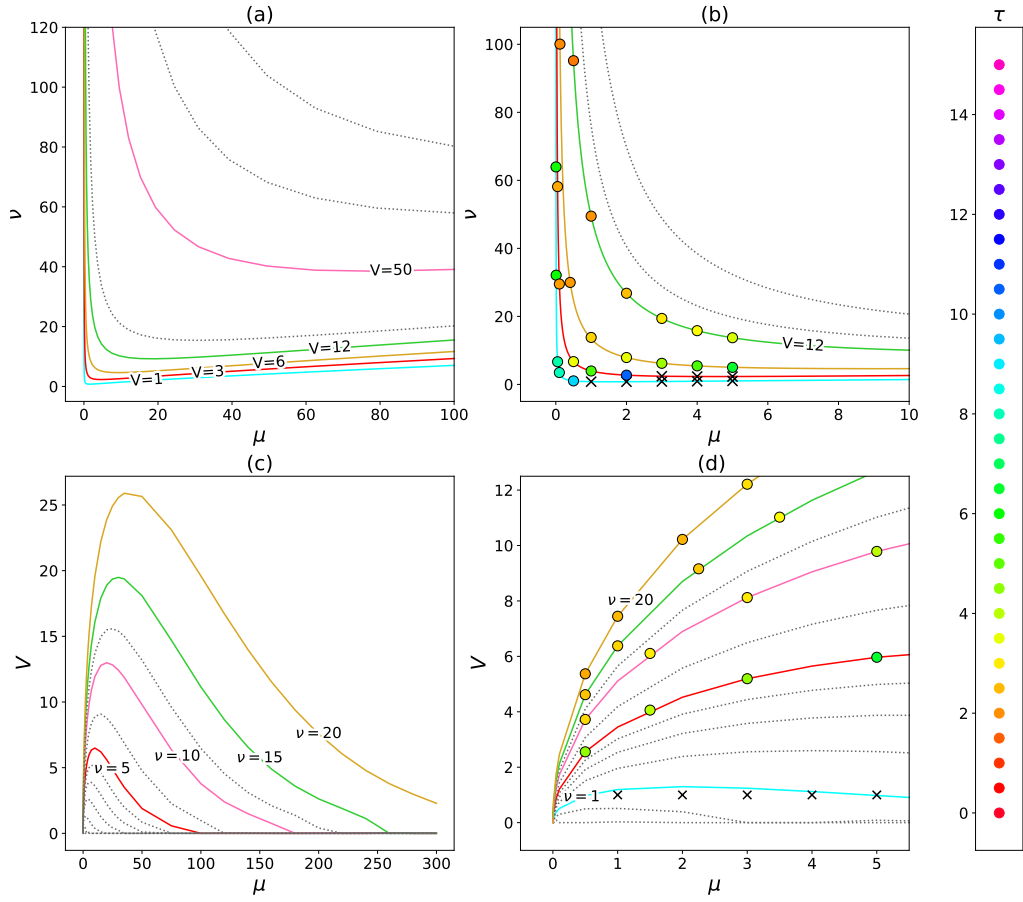


Figure 3.3: Visualizations of the four dimensional parameter space of  $V$  from solving Eq. (3.9),  $\tau$  from fitting the model OCF, along with model parameters  $\nu$  and  $\mu$ . (a) and (b) depict constant velocity  $V$  contours, while in (c) and (d) constant  $\nu$  contours are shown. Panels (b) and (d) show sub-regions of panels (a) and (c), respectively, with added individual points depicting values of  $\tau$  given by the color bar. Note that large  $\tau$  values (greater than 15) are colored black.

our model (solid lines). This overestimation is larger for the two larger values of  $\mu$  that correspond to weaker self-avoidant memory, as shown in Fig. 3.2f. Alternatively, using  $V$  from Eq. (3.9),  $\tau$  is determined by fitting the long time MSD to Eq. (3.11). These values of  $\tau$ , plotted as the dashed lines in Fig. 3.2d, substantially underestimate the decorrelation time scale of our model, also shown by the corresponding dashed lines of exponential decay in Fig. 3.2c. Although the form of exponential decay of the orientational persistence is consistent with ABP and quantifiable by  $\tau$ , it alone is not enough to predict the enhanced diffusion of our model. There are additional effects of self-avoidant memory beyond the persistence memory, which is the only memory present in ABP.

At constant velocity, the effect of increasing self-avoidant memory (decreasing  $\mu$ ) is seen in Fig. 3.2a. The MSDs with smaller  $\mu$  in both cases depart from the ballistic regime earlier, and thus exhibit less enhanced diffusion. This corresponds to Fig. 3.2d where for smaller  $\mu$  the OCF decays more rapidly as measured by a smaller value of  $\tau$ . This is further illustrated in Fig. 3.3b, showing a decrease of  $\tau$  with decreasing  $\mu$  along contours of constant velocity. For fixed  $\mu$ ,  $\tau$  increases with decreasing  $\nu$  (although velocity decreases). Thus we see that one effect of self-avoidant memory as it is present in our model is to decrease orientational persistence: swimmers with high memory experience weak orientational persistence and vice versa.

A more exotic effect of self-avoidant memory is shown by the trajectories in Fig. 3.4 and provides a plausible explanation for the surprising fact that ABP overestimates the enhanced diffusion of the model. To avoid crossing their own self history, paths turn back on themselves and continue turning inward, becoming caged for a while before enough diffusion has occurred for them to leave this self-created trap. This transient self-trapping perhaps explains the reduced enhanced diffusion as compared to ABP with equivalent orientation persistence time. Self-trapping has been studied in autophoretic systems like that which we model, although it has only been found in chemoattractant experimental systems [41, 75] and model systems [62, 25, 38] with self-attracting memory. It will be interesting to find out whether self-avoidant experimental systems like that in [44] show similar self-trapping.

### 3.4 Limiting Behavior

Since the relevant experimental systems are well described by ABP and we can explicitly tune memory in our model, we anticipated that we could find a parameter regime (low memory, high  $\mu$ ) in which the enhanced diffusion produced by our model was well described by ABP. However, as discussed at the end of Sec. 3.2, the

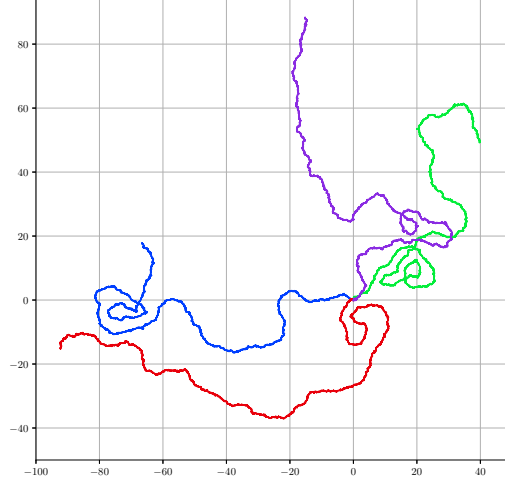


Figure 3.4: With  $\mu = 0.01$  and  $V = 6$ , four sample paths are shown which illustrate the caging of enhanced diffusion experienced due to high memory.

limiting behavior of systems (3.2) and (3.3) as  $\mu \rightarrow \infty$  is simple Brownian motion, indicating that enhanced diffusion with low self-avoidant memory may not be possible. We revisit this limit with further simulations in light of the emergent parameters  $V$  and  $\tau$ , considering both  $\nu \rightarrow \infty$  with  $V$  fixed as well as dynamic  $V$  with  $\nu$  fixed. Additionally, we investigate the high memory limit ( $\mu \rightarrow 0$ ), and find it consistent with Eq. (3.5) describing classical Brownian motion with (unfortunately) no further memory effects to investigate.

From Figs. 3.3a and b, we can consider the limit most likely to produce enhanced diffusion consistent with ABP: removing the memory via the limit  $\mu \rightarrow \infty$  while keeping the particle at constant velocity by fixing  $V$ . Visually we observe that as  $\mu \rightarrow \infty$ , the contours of  $V$  become flatter, reproducing the behavior seen in Fig. 3.1 which shows that ballistic motion is sensitive to changes in the gradient response  $\nu$ . Moreover, remaining on one  $V$  contour requires  $\nu \rightarrow \infty$  much slower than  $\mu \rightarrow \infty$ . To investigate the enhanced diffusion in this limit, we look at  $\tau$  in Fig. 3.3b. Following a  $V$  contour as  $\mu \rightarrow \infty$  results in an increase in  $\tau$  corresponding to longer orientational persistence (or less change in direction).

As a result of decreasing the self-avoidant memory timescale ( $\mu \rightarrow \infty$  while maintaining a constant velocity  $V$ ), we find that both the past history of the trajectory and the Brownian noise become less important in influencing the future location of the trajectory. Furthermore, with the addition of an increased gradient response by taking  $\nu \rightarrow \infty$  (as required to keep the particle at constant  $V$ ), the deterministic gradient response force in Eq. (3.2) dominates the Brownian noise, and consecutive steps become more correlated. This increases the persistence time  $\tau$ , and the trajectories approach purely ballistic motion with no enhanced diffusion at observable finite times.

Figs. 3.3a and b also allow for considering infinite memory ( $\mu \rightarrow 0$ ), again with constant velocity. In Fig. 3.3a, we see that the gradient response required (given by the size of  $\nu$ ) to keep the particle moving at constant velocity  $V$  rapidly blows up to  $\infty$ . This is largely unsurprising as the prefactor on the deterministic term in Eq. (3.3) contains the product  $\mu\nu$ ; taking  $\mu \rightarrow 0$  while keeping this integral response factor relatively constant would necessitate  $\nu \rightarrow \infty$ . Fig. 3.3b shows a corresponding decrease in  $\tau$ , limiting towards pure diffusion. Returning to Eq. (3.2a), as  $\mu \rightarrow 0$  both the diffusion and the source term go to zero, thus the concentration field would remain fixed in time. If this initial concentration field was constant, then the particle would have no gradient to respond to and therefore only undergo pure Brownian motion in this infinite memory regime, corresponding to  $\tau = 0$ . This suggests that rather than trying to start at finite  $\mu$  and witness the effects of self-avoidant memory fade as  $\mu \rightarrow \infty$ , as this model was set up to do, future work should perhaps remove  $\mu$  from the source term in Eq. (3.2a) and start at  $\mu = 0$  to witness the effects of self-avoidant memory fade as  $\mu$  increases away from zero.

The limiting ballistic motion when taking both  $\mu$  and  $\nu$  to infinity is in contrast to the limiting Brownian motion behavior of Eq. (3.3) as  $\mu \rightarrow \infty$  while keeping  $\nu$  fixed. By following contours of  $\nu$  in Fig. 3.3c, we see that the velocity first increases with  $\mu$  and then decreases, approaching zero velocity as  $\mu \rightarrow \infty$ , which is consistent with the trend shown in Fig. 3.1b. It is interesting to observe in Fig. 3.3d that  $\tau$  appears to be relatively static along the contours of  $\nu$ . Note that these values of  $\tau$  were mainly computed at points to the left of the maximum velocity of the fixed  $\nu$  contours. Figure 3.1b indicated that  $\tau$  decreases with increasing  $\mu$  and fixed  $\nu$ . When we also consider a decrease in velocity, the trajectories can be assumed to approach Brownian motion.

In summary, increasing  $\mu$  to decrease the effects of memory either results in increasing  $\tau$  (by fixing  $V$ ) and therefore creating straighter trajectories that do not display enhanced diffusion in the MSD over the timescale of the simulation, or in decreasing  $V$  to zero (by fixing  $\nu$ ) which results in a purely diffusive MSD. The memory is responsible for both the ballistic motion measured by velocity  $V$  and the effective rotational diffusion measured by orientational persistence time  $\tau$ , so naturally follows that these effects are both lost with increasing  $\mu$ . If the concentration field diffuses infinitely fast by taking  $\mu \rightarrow \infty$  with  $\nu$  fixed, we lose deterministic motion since the gradient of the concentration field is always zero at the particle's center and radially isotropic around the particle, thus the net force acting on the particle is always zero. The effective angular diffusion is lost when taking  $\mu \rightarrow \infty$  with  $V$  fixed because this requires large  $\nu$  such that the immediate deterministic forces overwhelm the noise and any past history, and so reduce the MSD to

almost exclusively ballistic motion. Thus, incorporation of self-avoidant memory is not simply an addendum to the active Brownian model that can be removed without consequence; by its complex interactions with the enhanced diffusion we see that it makes for a categorically unique model.

### 3.5 Conclusions

We have analyzed the self-avoidant memory effects of a model coupling an active swimmer and an environmental chemical field. Like the experimental system it was inspired by, it can exhibit ABP-like behavior with the MSD having both a ballistic and a long-time enhanced diffusion regime. With an analytical formula for the velocity,  $V$ , we faithfully reproduced the ballistic regime. The enhanced diffusion in our model is a result of both angular persistence and the self-avoidant memory, whereas ABP only includes orientational persistence. We found that numerically computing the orientation decorrelation (or persistence) time,  $\tau$ , enhanced diffusion predicted by ABP overestimates the enhanced diffusion in our model. Thus, our proposed model did not faithfully capture the dynamics of the experimental system at long time scales in the same way that ABP did. (Further investigation will be needed to determine if this difference is due to parameter values, modeling choices like using thermal noise and the diffusive scaling to the source term, or the absence of hydrodynamic effects.) Instead, we discovered that the self-avoidant memory in our model led to transient self-trapping that suppressed the enhanced diffusion. This self-trapping has, to date, been suggested to occur only in self-attracting experimental systems [41, 75] and computational models [62, 25, 38]. Further investigation will be needed to determine if self-trapping is a unique feature of this model, or can occur in other self-avoidant systems.

Through these investigations, we kept the noise parameter  $\epsilon$  fixed, while changing the gradient response parameter  $\nu$  and the diffusion  $\mu$  to find that both latter parameters control the implicit parameters  $V$  and  $\tau$ . With only two control parameters, we were unable to independently tune each timescale of behavior: the velocity  $V$ , the memory timescale  $\mu^{-1}$ , and the angular persistence timescale  $\tau$ . Taking  $\mu \rightarrow \infty$  to remove memory effects, we either arrived at simple Brownian motion by fixing  $\nu$  or purely ballistic motion by fixing  $V$  and allowing  $\nu \rightarrow \infty$ ; the memory is responsible for both the ballistic motion and the effective rotational diffusion. Taking  $\mu \rightarrow 0$ , we again arrive at simple Brownian motion, having removed all self-avoidant memory with our choice of scaling the source term in the concentration field by  $\mu$ . We thereby identified an intermediate regime of  $\mu$  for which enhanced diffusion is present on a finite timescale, but at a lower

magnitude than expected for ABP with equivalent angular persistence. This regime will be used in future work to study self-avoidant memory effects in many-particle simulations, investigating motility induced phase separation and associated dynamic pattern formation, which is commonly observed in active systems with particles that are repulsive to one another.

## CHAPTER 4

### Unpacking Emergent Behaviors of Self-Avoidant Swimming Droplets with Curvature Statistics

#### 4.1 Introduction

In the previous chapter, we discovered discrepant displacement statistics between self-avoidant swimming droplet paths and active Brownian paths (which are commonly used to model active systems) at long timescales. We proposed that the reduced effective diffusion at long timescales found in [15] is produced by extremely high curvature in the self-avoidant particle path data, which is due to high levels of self-interaction. Despite trying to “get away from themselves”, self-avoidant particles may become trapped in their own trails, and then continue turning inward to avoid crossing their past history with the result that they move smaller distances on long timescales.

If this self-trapping occurs (and sometimes it does not), it manifests quantitatively as regions of high path curvature. In path data, curvature occurs when the particle makes consecutive turns in the same direction. To start, we considered the more general “average turning/reorientation timescale”,  $\tau$ , in [15] to measure when the average movement direction becomes decorrelated. In other words,  $\tau$  can estimate how often a particle turns in *any* direction, which is less specific than we find we need; decorrelated motion can be a product of many changes in direction, but not *necessarily* consecutive changes in the same direction, as in self-trapping. (Linearly correlated motion corresponds to “straight” path data.) In the ABP model,  $\tau$  controls the reorientation timescale directly since it is the rotational noise strength in equations (ABP model), whereas in the self-avoidant model,  $\tau$  can only be estimated by fitting the decay of linear correlations in orientations within the path data, which is assumed to have the form  $e^{-\frac{t}{2\tau}}$  [33]. Ultimately, it is our goal to estimate correlations in curvature, rather than merely correlations in direction.

The results of this exercise are seen in Fig. 4.1. Despite matching  $\tau$  fitted from the self-avoidant velocity autocorrelation function (VCF, also called the orientation correlation function, or OCF) shown in Fig. 2.1c and with the choice of  $\tau = 2.3$  in the ABP model for the paths in Fig. 2.1a, there is noticeably more curvature in the self-avoidant paths. We argue that different turning motivations is the source of this difference. In

the ABP model, curvature is a result of decorrelations in the direction, however, in our model, curvature is *also* a result of self-trapping. When our particles reorient inward (randomly) the deterministic gradient force overrides the effects of the random noise and the particle continues the inward spiral by then making consecutive turns in the same direction. Since  $\tau$  is constructed to control linear decorrelations in direction, it works well to characterize the ABP paths. In contrast, it fails to adequately characterize the observed self-trapping in our model paths since the self-trapping is not a product of decorrelations. In fact, self-trapping (and resulting curvature) is a mesoscale *emergent* effect of self-avoidant memory, since it arises organically as the system evolves, but is not a result of any individual model component. We conclude that a single reorientation timescale which captures the decay of *linear* correlations between consecutive orientations is not a good classifier of trajectories exhibiting curvature which arises from self-avoidant memory.

## 4.2 Calculating Curvature Directly

We determined previously the decay of linear correlations in direction measured by  $\tau$  was insufficient to describe self-trapping curvature since the curvature was not a product of random directional changes. In our model, the true *source* of the curvature is the gradient forces experienced by the particle, which temporarily override the directional changes that may be induced by the additive noise. In this section, we will estimate the curvature of paths directly; this data will be the foundation for us to build a better tool for understanding how the curvature can be used to quantify the emergence of a self-avoidant memory response. To estimate the curvature, we borrow the multi-scale straightness index (MSSI), a tool from mathematical ecology that allows us to choose what level of granularity we want to see in the trajectory data [55]. The MSSI is defined to be the ratio of linear displacement to arc length for a coarse-grained path sample of frequency  $g$  in a moving window of size  $w$  over the experimental time frame (see Fig. 4.2). MSSI values close to 1 indicate very straight motion (low curvature) and values close to 0 indicate high curvature. In [55], this tool is used on possum trajectory data to classify possum behavior as either foraging (high curvature) or traveling (straight).

Inspection of the trajectories in Fig. 4.1 shows that the self trapping occurs on an intermediate spatial scale, therefore, we tune the MSSI to the mesoscale at which self-trapping occurs to capture the temporally changing path curvature. We choose  $g$  to smooth out small-timescale diffusion and  $W$  on the ballistic timescale, on the order of  $V$ . Fig. 4.3 shows that the average straightness of model paths with standard

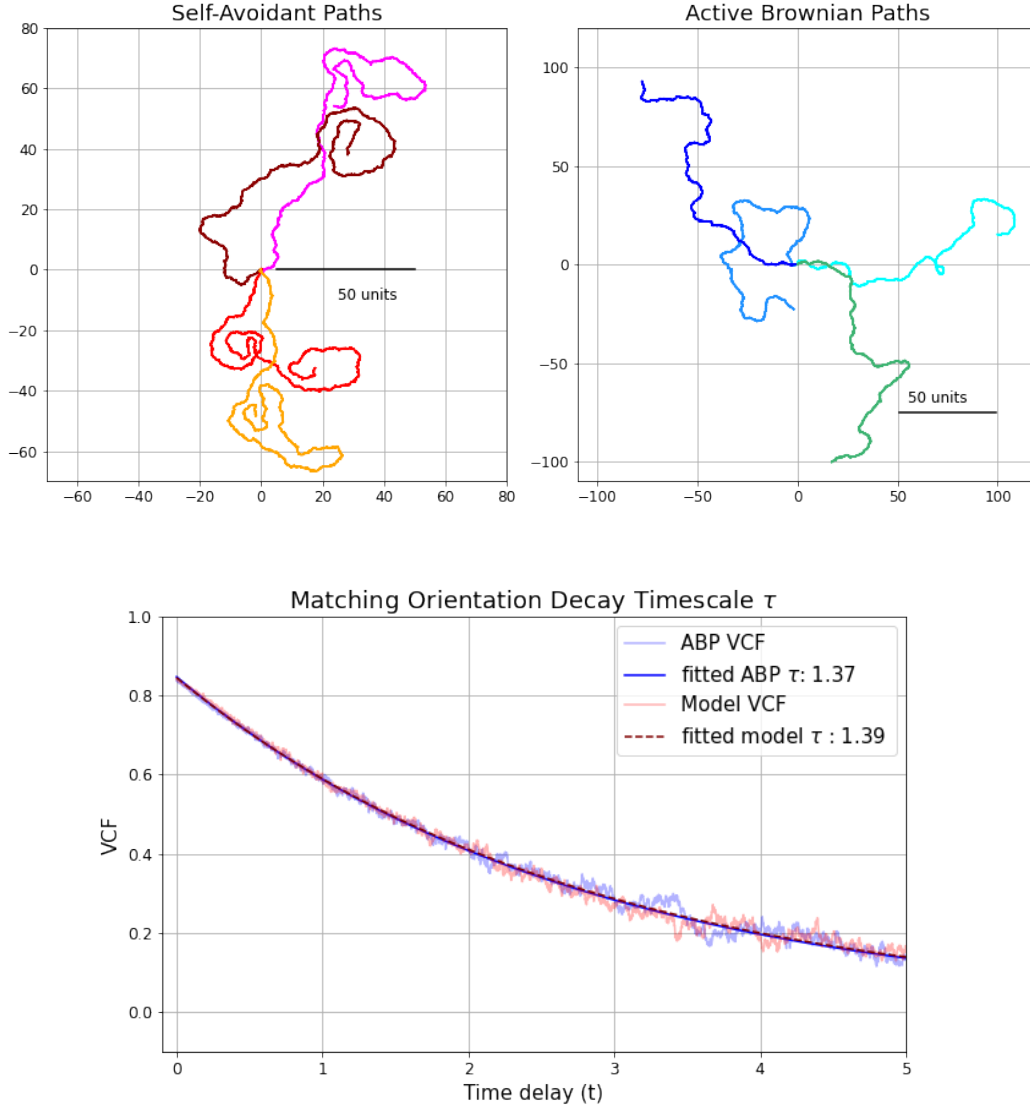


Figure 4.1: Similar turning times generate different curvature features in different models. (a) Self-avoidant paths with average turning time  $\tau = 1.4$  are highly tortuous. (b) ABP paths with the same theoretical turning time  $\tau = 1.4$ . (c) Both models recover the same VCF statistics (fitted  $\tau$ ), despite obviously different path features, including differences in overall distance covered. To compute the instantaneous velocity from the path data, we compute the displacement between positions which are  $N$  timesteps apart. We choose this calculation timescale  $N$  to be the timescale over which the expected displacement is approximately 25 times larger than the expected displacement due to the white noise alone:  $NVdt \approx 25 \cdot \epsilon\sqrt{dt}$ . For  $dt = 0.00125$ ,  $\epsilon = \sqrt{0.75}$ , and  $V = 6$  we find that  $N = 102$ .

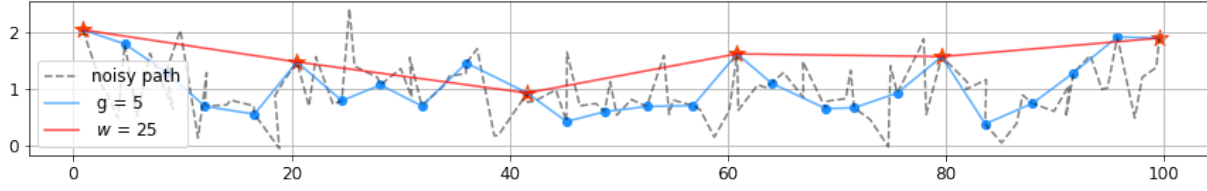


Figure 4.2: Sampling scheme for noisy data. Linear displacement is computed between the endpoints of each moving window of size  $w$  and arc length computed along the path sampled from  $w$  at frequency  $g$ .

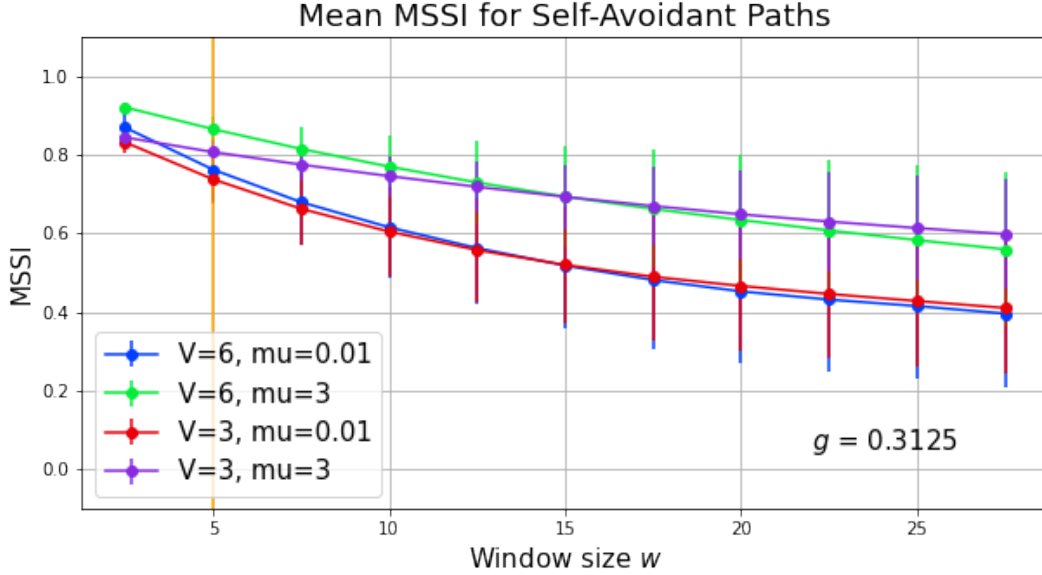


Figure 4.3: Average MSSI values of self-avoidant paths changes with window size. For all self-avoidant parameters ( $V$  and  $\mu$ ), the straightness decreases as the window size  $W$  increases (coarse-graining). The gold line indicates the window size at which the MSSI is computed for the plots and calculations in the remainder of the chapter. Window size  $w$  and granularity  $g$  are in units of nondimensional time.

errors for several parameter regimes. We see that the straightness decreases as a function of window size  $w$ . Increasing  $w$  corresponds to increasing the spatial scale across which the curvature is estimated.

Our claim is that the curvature is an emergent mesoscale effect of the self-avoidant memory; more specifically, we assert that randomly developing past curvatures build up and influence future curvatures. To capture this evolution, we represent the MSSI in a time-dependent manner and construct a time-series of MSSI values which functions as a time-evolving order parameter for the curvature on a path-by-path basis. By considering individual trajectories we find that the MSSI captures the curvature well, as it varies throughout the experiment. In Fig. 4.4 we have selected two paths with the same model parameters, but which have very different path features. The one on the left is more similar to active Brownian motion (no

memory) because it does not trap itself, and the one on the right exhibits the self-trapping which we believe is an emergent feature of the self-avoidant memory. The curvature differences in both are reflected in the values of the straightness index time series (to the right of each path).

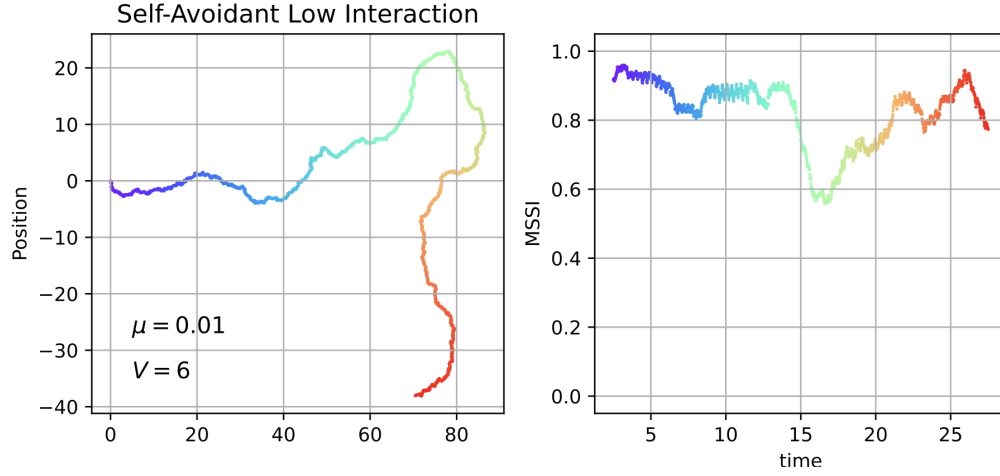
One of the reasons why average turning timescale  $\tau$  failed to adequately explain the curvature, was the fact that  $\tau$  was computed from the ensemble average VCF, which tended to average out the self-trapping behavior, which occurs less frequently compared to more generic swimming-dominant trajectories. While self-trapping was evident in some paths, the overall distribution of curvature shows that the parameter regime we are working in produces straighter paths, on average. We see this in the histograms of Fig. 4.5, which carry most of the empirical probability mass near 1, although the empirical cumulative density function in Fig. 4.5 shows that true density of the straightness values of the self-avoidant paths do appear to have a heavier tail compared to the active Brownian paths. We conclude that self-interaction which are associated to high curvature (low straightness) occur much less frequently than generally ballistic motion, even though these self-interaction events affect the ensemble statistics of the displacement data which we explored in Chapter 3.

### 4.3 Using Mutual Information to Estimate Nonlinear Correlations in Path Curvature

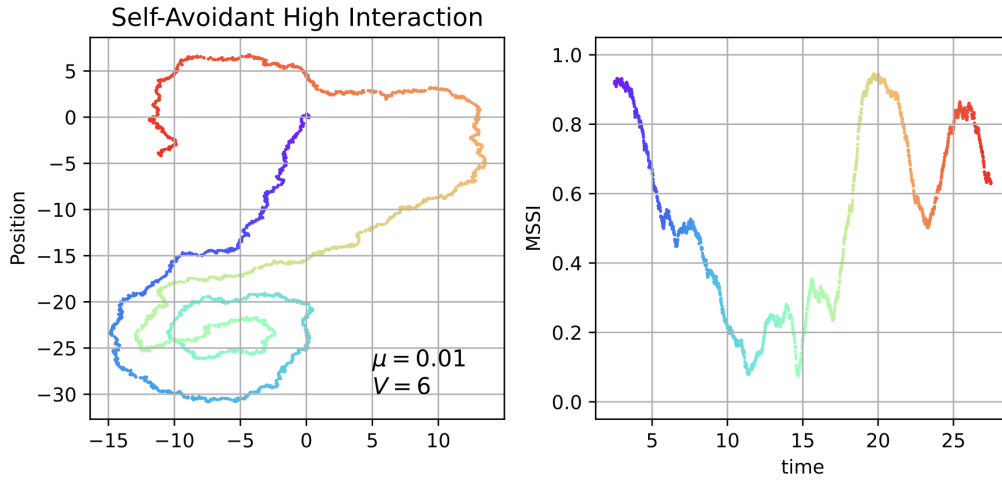
In the previous section, we presented a direct way of quantifying curvature with the MSSSI, but this does not directly connect the curvature to the self-avoidant memory. We demonstrated that using the decay timescale of *linear* correlations in the orientations of our model paths was insufficient to explain the self-trapping, which is a complex effect of the self-avoidant memory and not merely caused by random reorientations. We propose that this can be resolved by:

1. Using the curvature data itself rather a related approximation ( $\tau$ ), and
2. Considering the temporal structure of *nonlinear* correlations, rather than linear ones.

We will use mutual information to capture nonlinear curvature correlations of a self-avoidant particle with its own past history. Mutual information addresses the question: *if we know the curvature of a single path at past time  $t$ , how much information do we gain about the curvature at future time  $t + T$ ? (and vice versa)*. If the answer is “a lot”, then we can say we have shown that the curvature at current times is a response to the curvature at past times, and therefore, the curvature is capturing the self-avoidant response. We will show



(a) active Brownian-like model path



(b) model path with self trapping

Figure 4.4: Using a moving window, we can compute the straightness of a path at each point in time. In doing so, we can see that the straightness index captures the bendiness over time. We have selected two paths with the *same* model parameters, but which have very different path features. The top path (a), is more similar to active Brownian motion (no memory) because it does not trap itself, and the bottom path (b) exhibits the self trapping which we believe is an emergent feature of the self-avoidant memory. The curvature differences in both are reflected in the values of the straightness index time series (to the right of each path).

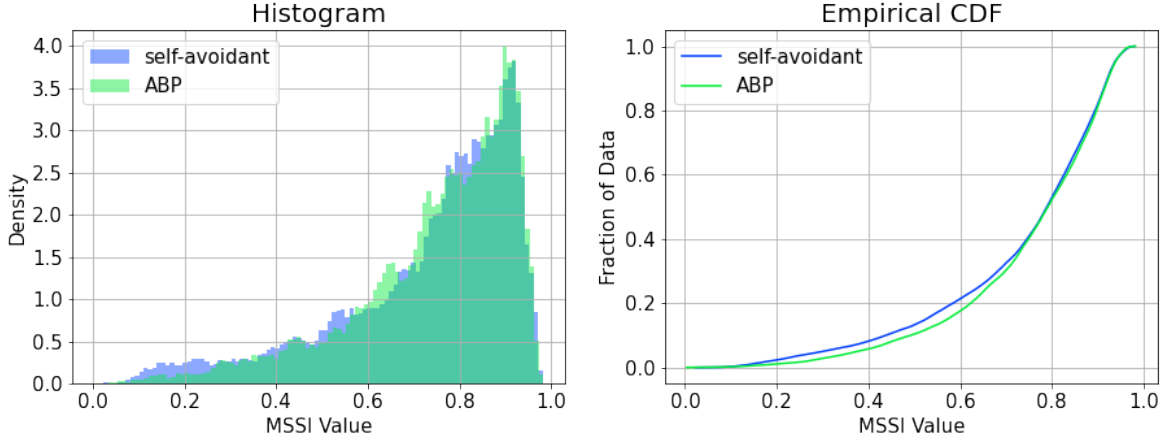


Figure 4.5: The ensemble histogram of MSSI values over all times illustrates the heavy bias toward “swimming-dominant” behaviors in the self-avoidant particles (the sharp peak near 1). (Self-avoidant paths computed with  $V = 6$  and  $\mu = 0.01$  and  $\tau$  fit to 1.389. ABP paths have matching  $\tau$  and  $V$ .) The empirical cumulative density functions show that the self-avoidant paths have a slightly higher density at lower straightness values.

that distributions of MSSI (curvature) of a single path at time  $t$  and time  $t + T$  are nonlinearly correlated, as a measure of self-interaction.

To compute the mutual information, we use the k-nearest neighbors algorithm described in section 2.1.2. In [32], the authors develop an unbiased statistical estimator for mutual information, which can be used on bivariate data of the form  $Z_k = \{(X_k, Y_k)\}$ , with  $k = 1, \dots, K$ . From  $Z_k$ , we select a random sample  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$ , with  $N \leq K$ , where this bivariate sample can be used under the assumption that each  $(X_n, Y_n)$ , is an independent realization from a stationary distribution. Since each path is generated independently, we have created an ensemble of independent agents whose straightness time series data  $\{S_1, S_2, \dots, S_N\}$  we estimate discretely throughout the experiment at times  $\{t_0, t_1, \dots, t_F\}$ . This ensemble is denoted as

$$S = \{\{(S_1(t_0), \dots, S_1(t_F))\}, \{(S_2(t_0), \dots, S_2(t_F))\}, \dots, \{(S_N(t_0), \dots, S_N(t_F))\}\}.$$

Therefore if the true distribution of  $S_n(t)$  is stationary, then sampling the MSSI variable  $S_n(t)$  corresponding to each of the  $N$  paths at some time  $t_i$  with  $i \in [0, 1, \dots, F]$  creates a random sample of  $N$  independent realizations,  $\{(S_1(t_i), S_1(t_i + T)), (S_2(t_i), S_2(t_i + T)), \dots, (S_N(t_i), S_N(t_i + T))\}$  for each time  $t_i$  with  $i \in [0, 1, \dots, F]$ .

In the case of the Ornstein-Uhlenbeck walkers, it is known that at long times, the positions of each walker sample from a stationary distribution. This is known because the evolution equation of an Ornstein-Uhlenbeck walker is known. However, we compute  $MI(S(t); S(t + T))$ , which is the mutual information of the ensemble of straightness index values; the representative dynamics (and associated stochastic differential equation) of the evolution of this random variable is unknown. Therefore, we provide qualitative evidence that the true distribution of  $S(t)$  is stationary to justify considering the mutual information of an individual path's straightness index with itself, which is  $MI(S_n(t); S_n(t + T))$ . First, we show in Fig. 4.6 that the ensemble mean and variance ( $\overline{\{S(t)\}}$  and  $Var(S(t))$ ) appears to be relatively constant as  $t$  changes; these constant descriptive statistics suggest that the underlying distribution of  $S(t)$  is unchanging. Secondly, consider that the Ornstein-Uhlenbeck walkers sample from a stationary distribution at long times  $t$  due to mean reverting tendency of the walkers, which functions as an effective potential well that the walkers are exploring. The effective potential well thereby reduces the effective support of the Ornstein-Uhlenbeck walkers. Similarly, the support of our variable,  $S(t)$ , is the closed interval  $[0, 1]$ , and at long times, the ensemble is more likely to reach a stationary distribution by thoroughly exploring this finite support.

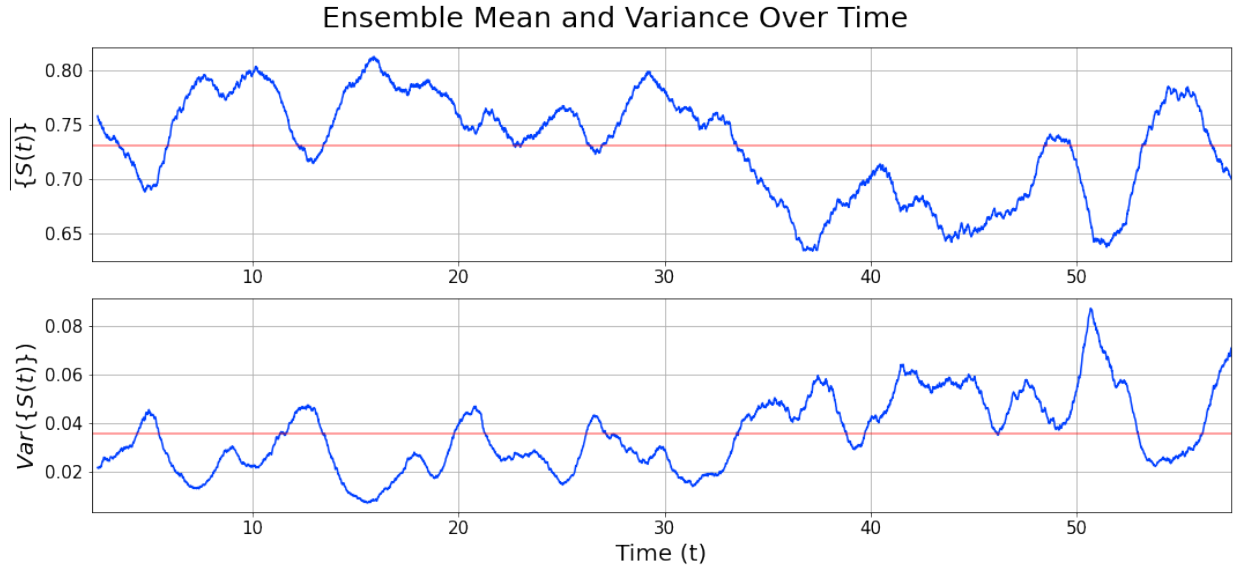


Figure 4.6: Sample mean and variance of  $\{S(t)\}$  over  $t$  appear to be fluctuating around constant values. Time averaged sample mean and variance are shown in red.

In Fig. 4.7, we illustrate a small ensemble of the straightness data of  $N = 5$  agents. We illustrate how we can construct a sample of independent realizations by taking each  $S_n(t = 5)$  (gold stars) and  $S_n(t + T = 5 + 1.5)$  (cyan circles). This generates a random sample of the ensemble at times  $t = 5$  and

$t + T = 6.5$  from which the mutual information can be estimated as  $MI(S(5); S(6.5))$ . By applying the estimator at each experimental time  $t_i$ , we can leverage the independence of each path to report the “true mutual information” of time-separated curvatures of the ensemble of  $N$  independent agents as it evolves throughout the experimental time frame (as we did with the Ornstein-Uhlenbeck walkers). By varying the time-delay  $T$ , we can assess whether looking back farther in time changes the mutual information of the system.

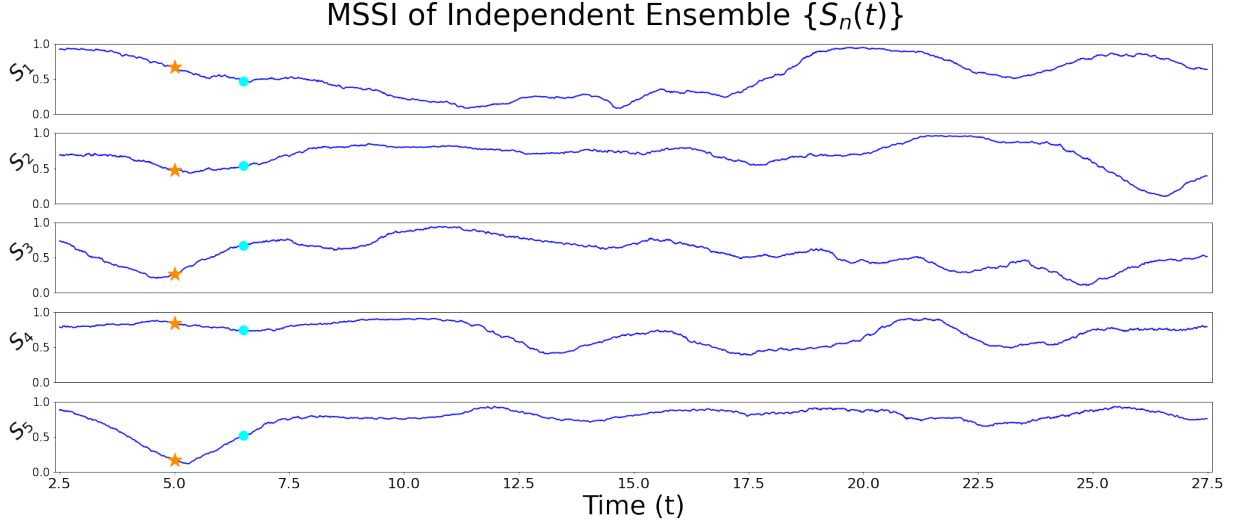


Figure 4.7: For each path  $S_n$ , we take the data at some time  $t_i$ , marked with an orange star, and a time-delayed sample at time  $t_i + T$ , marked with a cyan circle. For this illustration, we choose  $t_i = 5$  and  $T = 1.5$ . Since each  $\{S_n\}$  is independent, the sample set  $\{(S_1(5), S_1(6.5)), \dots, (S_5(5), S_5(6.5))\}$  contains independent realizations of the random variable  $S$ .

Unlike the Ornstein-Uhlenbeck walkers, whose mutual information gradually converges to a constant value (Fig. 2.8), the self-avoidant straightness ensemble mutual information appears to be fluctuating around a constant value throughout the entire experimental time frame, which we show in Fig. 4.8. This suggests that the true time-delayed mutual information of the independent ensemble  $S$  is constant in time. We see that the ensemble mutual information follows this trend for all time-delays  $T$ , but looking back farther in time (which corresponds to increasing  $T$ ) lowers the mutual information until it is approximately zero. We have shown that the ensemble curvature at time  $t_i$  contains information about the ensemble curvature at future time  $t_i + T$ , and these nonlinear correlations decrease as  $T$  increases. Although we assumed (and presented evidence for) stationarity of  $S(t)$ , we do not have any information about the actual distribution of  $S(t)$  itself.

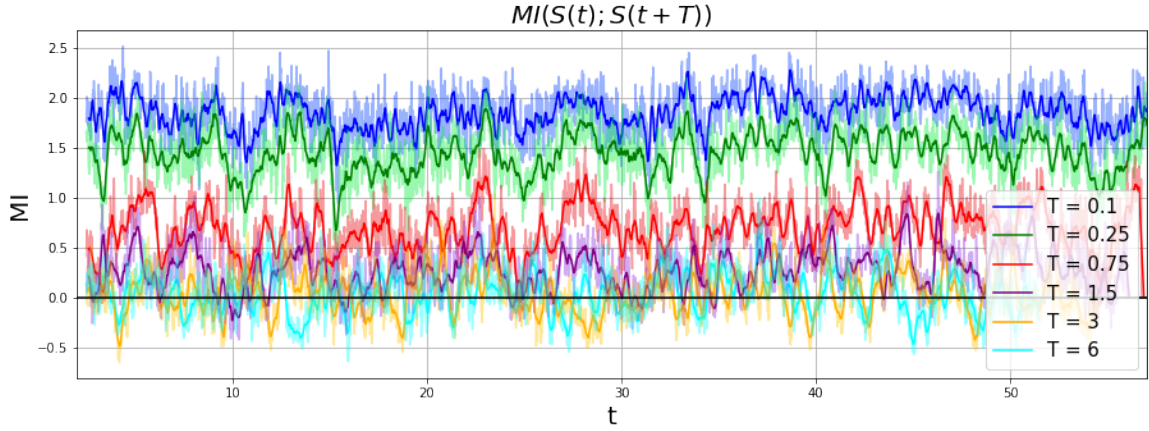


Figure 4.8: The independent ensemble mutual information appears to fluctuate about a static mean throughout the experimental time frame. Dark lines indicate a 20 second moving average.

#### 4.4 Time Delayed Self Mutual Information of a Single Agent

We now implement our mutual information estimator pipeline to estimate the mutual information between straightness data of a *single path with itself* at later times, which we denote as  $MI(S_n(t), S_n(t + T))$ . Recall that both  $S_n(t_i)$  and  $S_n(t_j)$  are realizations of the continuous stochastic process  $S_n$ , and therefore we expect that  $S_n(t_i)$  and  $S_n(t_j)$  will be correlated to some extent, especially if  $t_i$  and  $t_j$  are very close in time. (The same is true of  $S_n(t_i + T)$  and  $S_n(t_j + T)$ ). As before in the ensemble case, we are challenged to ensure that consecutive samples  $(S_n(t_i), S_n(t_i + T))$  and  $(S_n(t_{i+1}), S_n(t_{i+1} + T))$  are independent samples of the stochastic process  $S_n$  with respect to the time separation *between pairs*, which is  $t_{i+1} - t_i$ .

We develop the solution to this problem which we developed in Chapter 2 by introducing a parameter  $W$  which we will enforce as the average window of separation between consecutive sample times  $t_i$  and  $t_{i+1}$ . In Fig. 4.9 we illustrate implementation of the separation window  $W$  using the straightness data of a single path  $S_n$  from the ensemble (shown in Fig. 4.7). This path is then sampled at discrete times  $\{t_i\}$  with average separation  $W$  and again at delayed times  $\{t_i + T\}$  to generate the random sample:  $\{(S_n(t_0), S_n(t_0 + T)), S_n(t_1), S_n(t_1 + T)), \dots, S_n(t_f), S_n(t_f + T)), \}$ . The randomly selected sample times  $[t_0, \dots, t_f]$  have the property that  $\langle t_{i+1} - t_i \rangle_i = \langle W_i \rangle_i \approx W$ .

We want  $W$  to be large enough to overlook dynamical correlations, but small enough to capture a signal. To justify the choice of  $W$ , we might naturally choose  $W$  to be the timescale on which the autocorrelation of the MSSSI time-series decays below some threshold value. In Fig. 4.10, we find that the linear correlations in

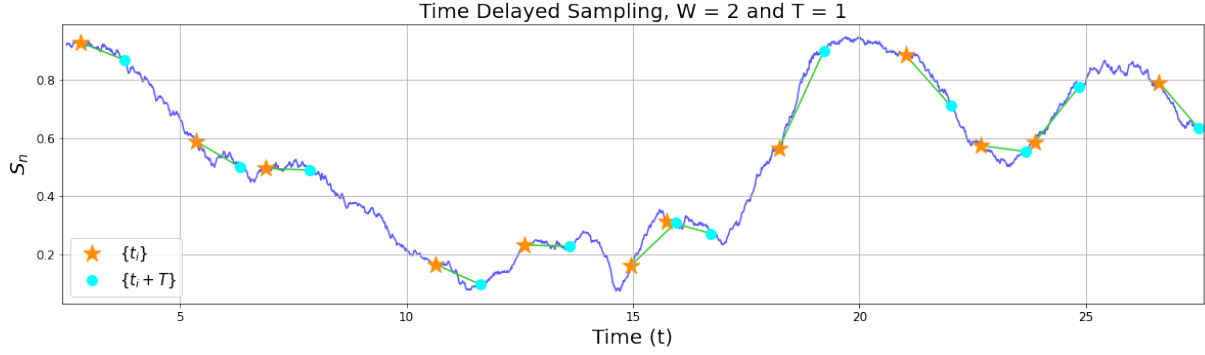


Figure 4.9: Gold initial samples with times  $\{t_i\}$  must be separated by windows  $W_i$  satisfying  $\langle W_i \rangle_i \approx W$ . The time delay  $T = 1$  is then used to select the time-delayed curvatures (cyan samples).

the MSSI time-series are very long lived and oscillate, making this an unsuitable strategy. In practice, we suggest that sufficiently large  $W$  is  $W$  such that the MI curve decays to zero.

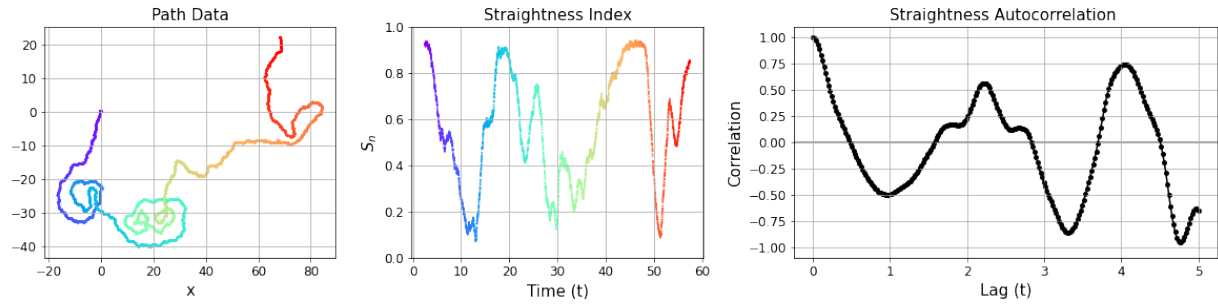


Figure 4.10: The autocorrelation function of the straightness of path data has no useful information to aid in proper choice of  $W$ .

## 4.5 Individual Path Analysis

Between individual paths, we aim to show that the mutual information can quantify, and then distinguish, the visually observable differences in path features. In the top row of Fig. 4.11, we consider three paths generated by the same model parameters ( $\mu = 0.01$  and  $V = 6$ ) which are selected for their individual characteristics. Path *A* appears the most similar to active Brownian motion, in which the high velocity  $V$  produces a trajectory that has a substantial ballistic component. Path *B* has a similar ballistic tendency but is shortly trapped once. Finally, path *C* is trapped substantially.

In the bottom row of Fig. 4.11, we show the mutual information decay curves for each path for window sizes  $W = 1, 2, 4$ . For all window sizes, the time-delayed self mutual information  $MI(S_n(t); S_n(t + T))$

decreases as a function of the time delay  $T$  in what appears to be an exponential decay. Recall that our particle model is designed to experience the influence of *all* past times as it moves to the next location. Since the past-history dependent memory effect of our model is encoded as an exponentially decaying memory kernel (see Eq. 3.3), we expect that more recent times will have a larger effect than locations farther in the past. In Fig. 4.11, we confirm that the mutual information between curvatures follow this same pattern; as the particle “looks back farther in time” and  $T$  increases, the curvatures become less correlated.

Although the overall decay structure is consistent across window sizes, we observe some differences. For a window size of  $W = 1$  (blue), the mutual information does not appear to decay convincingly to zero for any path on the timescales we have considered. The decay curve for  $W = 1$  appears to either decay to zero on a time delay  $T$  greater than what we consider, or not at all. This suggests that the separation window  $W$  is insufficient to suppress the dynamical correlations. The window size  $W = 2$  is an improvement, and the curves corresponding to  $W = 4$  all appear to decay to zero within the time frame that we consider. (Ideally, we might also test windows larger than  $W = 4$ , however, the length of our paths limits the number of samples we can achieve with this separation. For a total path length of 60s, a window of  $W = 4$  yields a maximum of  $N = 15$  samples. While the Kraskov algorithm is suitable for low sample sizes, we do not want to have unnecessarily small  $N$ .) On this largest window size,  $W = 4$ , we notice that mutual information of the active Brownian-like path curvature of path  $A$  appears to decay to zero more quickly than the self-interacting paths  $B$  and  $C$ , suggesting a longer correlation that we associate to the presence of the self-trapping behavior which we believe is an emergent effect of the self-avoidant memory.

To compare the decay curves of the individual paths quantitatively, we can fit a decaying exponential,  $e^{-\frac{t}{\alpha}}$  to the mutual information decay curve and use  $\alpha$  as a measurement for the decay timescale of these nonlinear correlations. We will fit the curves corresponding to the largest window size  $W = 4$  to ensure that they decay to zero. In Fig. 4.12 we plot the fitted curve  $e^{-\frac{t}{\alpha}}$  (solid lines) over the mutual information data (scatter points). Colors in the figure correspond to the path colors in Fig. 4.11. We find that the self-trapped path (purple) has the longest decay timescale of approximately  $\alpha \sim 1.3$ . In contrast, the active Brownian-like path (red) has a much smaller decay timescale of approximately  $\alpha \sim 0.72$ . (The “in between” blue path has a correspondingly in between  $\alpha$ ).

Although each path is generated by using the same model parameters, it cannot be predetermined whether a path will experience self-trapping or not. As the particles descend the gradients created by their own oil expulsion, the path preference is to move ballistically (with noise) away from areas of high concentration.

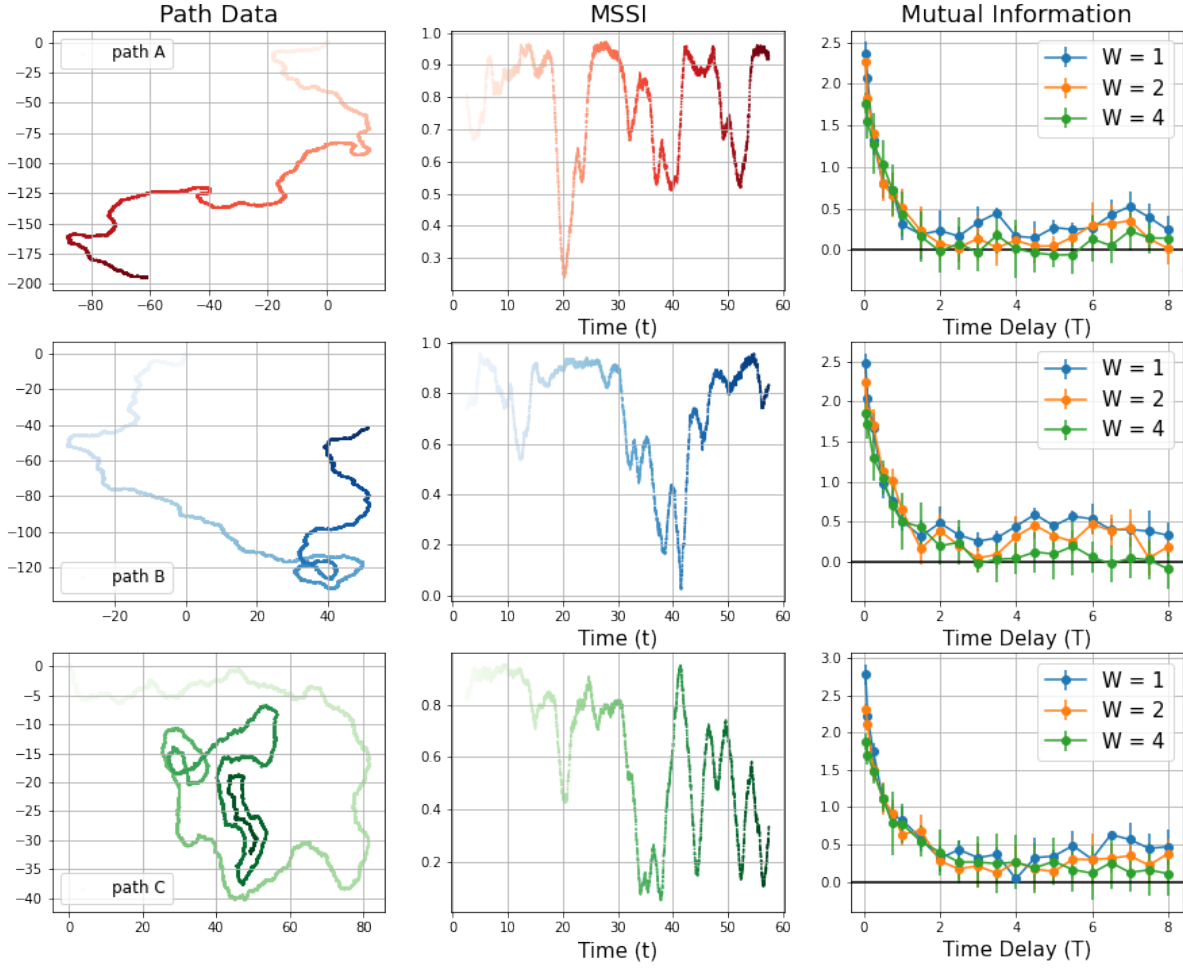


Figure 4.11: Three paths with different levels of self-interaction are shown (first column) with their corresponding straightness index time series data (middle column) and mutual information decay curves (left column). Mutual information decay curves with three window sizes,  $W = 1, 2, 4$  are shown as a function of the time delay,  $T$ . Reported values are the average of 20 repetitions with standard errors.

When the noise perturbs this motion such that the particle turns in on itself, the path preference changes from ballistic to inwardly curving motion in an attempt to avoid the self-created gradient trap. (This process continues until enough local gradient has built up to push the particle out of the trap and back into freer space.) Therefore, *randomly initiated curvature precedes periods of more predictable curvature*. We interpret the mutual information measurements of the curvature of a single particle as an indication of the strength of the self-avoidant memory response experienced by that particular particle over varying timescales. In doing so, we demonstrate that the more trapped paths have longer-lived correlations in the curvature, which is indicative of higher levels of self interaction via the self-avoidant memory response. We have shown that

the mutual information decay timescale  $\alpha$  can be used to statistically distinguish self-trapped paths from non-trapped paths, all of which have the same input parameters.

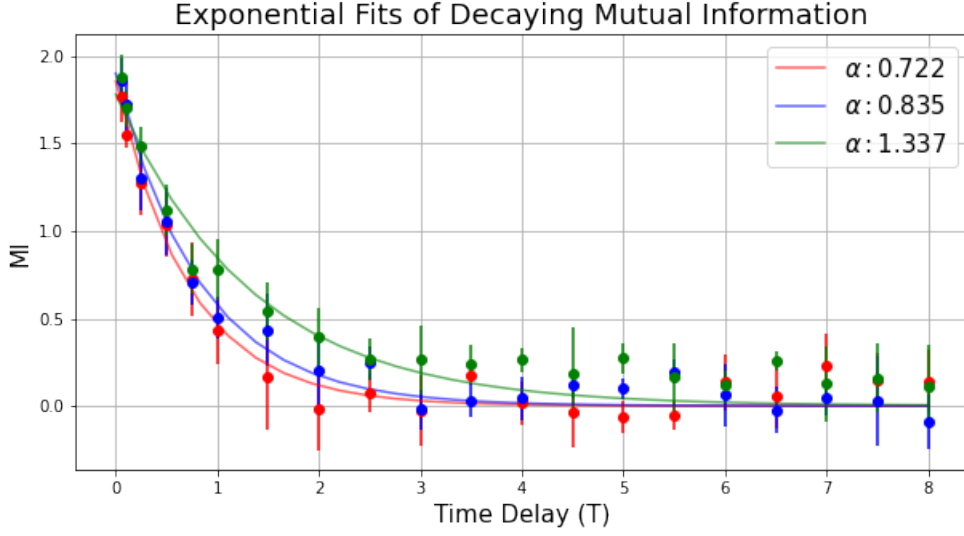


Figure 4.12: The mutual information decay curves (data points) of individual paths are fitted with a decaying exponential (solid lines). For more highly self-interacting paths (green), the decay timescale  $\alpha$  is higher as compared to more “active Brownian- like” paths (red). Colors correspond to paths featured in Fig. 4.11 and reported mutual information values are the average of 60 repetition with standard errors.

#### 4.6 Tuning Memory

We have discussed in the previous section that the self-avoidant memory response emerges more strongly in some paths compared to others, and we argued that this was unpredictable behavior since the genesis of self-trapping comes from the random noise (although the continuation of the self-trapping is a result of the deterministic gradient response). To explore the effects of changing the effective memory of the system itself, we recall the combined nondimensional form of our particle model:

$$d\mathbf{Y} = \frac{\pi}{2} \mu \nu \phi \left[ \int_0^t \left( e^{-\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(1 + \mu(t-s))}} (\mathbf{Y}(t) - \mathbf{Y}(s))(1 + \mu(t-s))^{-2} \right) ds \right] dt + \sqrt{\epsilon} d\mathbf{B}. \quad (4.1)$$

In this form, the integral term represents the particle mathematically “looking over all past times” to determine its next location. By increasing the lower bound from zero to  $\max(0, t - M)$  with  $M \geq 0$ , we can artificially restrict the amount of past history to affect the particle motion, and therefore reduce the effective memory

with the following particle path model:

$$d\mathbf{Y} = \frac{\pi}{2} \mu \nu \phi \left[ \int_{\max(0, t-M)}^t \left( e^{-\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(1+\mu(t-s))}} (\mathbf{Y}(t) - \mathbf{Y}(s))(1 + \mu(t-s))^{-2} \right) ds \right] dt + \sqrt{\epsilon} d\mathbf{B}. \quad (4.2)$$

Since the combined model in Eq. 4.1 is a non-Markovian process in which the state (position) at time  $t + \delta t$  depends on all past states via the integral term, implementation of effective memory timescale  $M$  is merely limiting the number of past states that the particle has access to from  $[0, t]$  to  $[\max(0, t - M), t]$ . (Unlike the combined model, the coupled model from Eq. 3.2 *can* be considered Markovian if the state space is expanded from the current location of the particle to both the location *and* current concentration field. Whereas implementing  $M$  in the combined system (Eq. 4.1) amounts to looking at fewer past states, there is no equivalent state space restriction for the Markovian coupled system (Eq. 3.2), which introduces a mathematical constraint to the implementation of effective memory timescale  $M$  in the coupled system.)

While the temporal effects of implementing effective memory timescale  $M$  are straightforward, the spatial effects require more explanation. In a physical sense, artificially restricting the amount of past history available for the particle to “see” by truncating the integral term in Eq. 4.1 to yield Eq. 4.2 is effectively removing all oil in the concentration field before time  $t - M$ . The effect of this is similar to evaporation or neutralization (which is possible experimentally), although both evaporation and neutralization are generally processes which occur smoothly in time rather than some amount of oil “disappearing” instantaneously. From the perspective of the particle, increasing the value of the diffusion coefficient  $\mu$  has the effect widening the tail of the particle’s trail, whereas implementing effective memory timescale  $M$  merely cuts off the tail beyond  $\max(0, t - M)$ . Introducing neutralization would have the effect of retaining the width controlled by  $\mu$ , but the peak will decay more quickly as the local concentration near the particle depletes more rapidly.

We illustrate these differences in Fig. 4.13, which shows a sketch of the concentration profile width emitted by a particle located at  $(0, 0)$  (black star) and travelling in a straight line along the x-axis under four different conditions. (The real local concentration profile is a mollified delta function centered at the particle location, therefore the particle does emit oil out in front of its own path. However, we reduce our sketch to only include the particle’s “wake” for clarity.) In blue we show the decay profile with some diffusion coefficient  $\mu$ . In red, we show the effects of increasing  $\mu$ —the faster diffusion both increases the decay rate (sharpening the slope away from the peak) and simultaneously widens the wake as emitted oil diffuses more rapidly into the environment. In cyan, the probable effects of neutralization can be seen in the faster decay

profile, although the width of the “wake” remains the same since the diffusivity  $\mu$  does not change. The depiction of implementing the effective memory timescale  $M$  is shown in the green curves, which abruptly end at the time  $t - M$ , which is the furthest past time that the particle has access to, as if the rest of the wake had simply disappeared. Ultimately, changes in  $\mu$  change the way that the trail diffuses in space behind the particle; increasing  $\mu$  spreads out the wake, thereby decreasing the concentration and reducing the memory strength. Depletion or neutralization does not affect the trail width; it reduces the concentration and therefore the memory by increasing the decay rate. Finally, truncation is not experimentally possible, but can be used mathematically to approximate neutralization.

We will be unsurprised to learn that changing the effective memory of our particle will change the velocity  $V$ . By assuming a steady state solution to Eq. 4.2 of the form  $\mathbf{Y} = \langle Vt, 0 \rangle$ , we can compute an expression which can be solved for the velocity,  $V$ :

$$1 = \frac{\pi \nu}{2 \mu} \phi \int_0^{\mu M} \frac{z}{(1+z)^2} \exp \left[ - \left( \frac{V}{\mu} \right)^2 \frac{z^2}{4(1+z)} \right] dz \quad (4.3)$$

after making the substitution  $z = \mu(t - s)$ . We fix  $\mu$  and  $\nu$  to the values which we know will result in self trapping, and solve for the corresponding  $V$  for selected values of  $M$ . In Fig. 4.14, we see how changing the value of  $M$  affects the velocity  $V$ . We see that at very small values of  $M$ , the velocity is zero until a threshold  $M$  is reached and the particle begins to swim. The exact threshold value of  $M$  can be found by setting  $V = 0$  in equation Eq. 4.3 and solving the following expression numerically for  $M$ :

$$1 = \frac{\pi \nu}{2 \mu} \phi \left( \frac{1}{\mu M + 1} + \log(\mu M + 1) - 1 \right) \quad (4.4)$$

The trend in Fig. 4.14 coincides with our intuition that looking back farther in time (increasing  $M$ ) is changing the effective gradient force experienced by the particle. Therefore, if  $M$  is small enough, the particle will not experience enough effective gradient force to move ballistically. After reaching the critical  $M$ , the velocity then rapidly converges to its true value of  $V = 6$ , corresponding to  $\mu = 0.011$  and  $\nu = 998.5$  (the parameters used to generate the paths of interest).

In Fig. 4.15, we see the effects of increasing the effective memory  $M$  for a single path. In all instances in the figure, the velocity  $V$  is beyond the swimming threshold depicted in Fig. 4.14. The full paths are shown in blue, and the path corresponding to  $[t - M, t]$  is shown in gold for  $t = 60$ . As  $M$  increases, we note that the

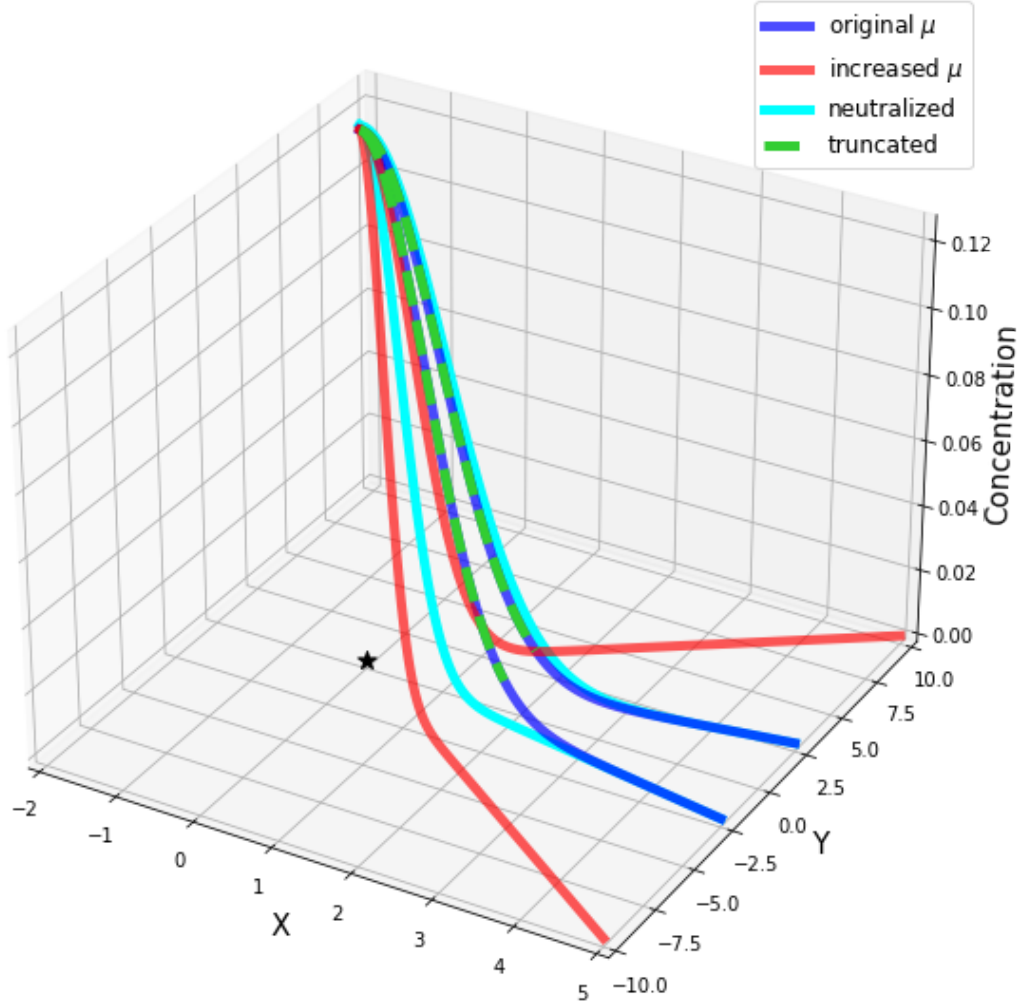


Figure 4.13: Crossection of the concentration profile of a particle at  $(0, 0)$  at time  $t$  shown under four different model conditions. In this artistic depiction, we suppose the particle is travelling through a free space with zero ambient concentration, thus the peak is created solely by the particle. We also neglect to illustrate the concentration “front” which would appear in front and on all sides of the particle as it travels- we only illustrate the train.

path covers progressively less space. We attribute this to the increased amount of gradient force perpendicular to the direction of motion as experienced by the particle. For increasing values of  $M$ , the gradient force of the larger gold path sections will affect the particle’s motion, causing it to reorient more often. Thus, the path with the smallest value of  $M$  appears “more ballistic” since it experiences mostly gradient forces parallel to the current direction of motion.

We now will explore how changing the effective memory  $M$  affects the mutual information of the path curvature  $S$ . For each value of  $M$ , we will compute a new ensemble of paths, where each ensemble has the

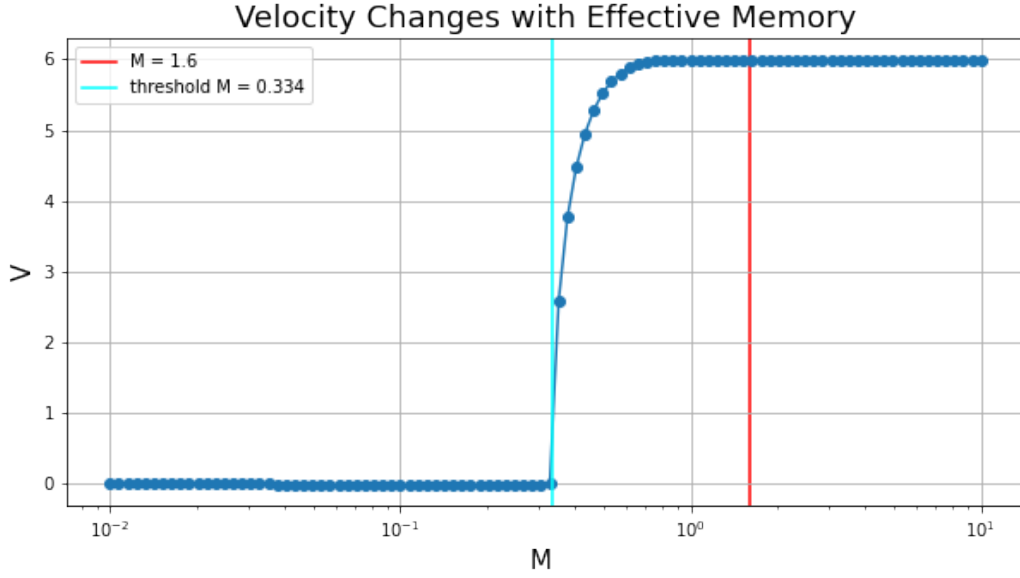


Figure 4.14: At small values  $M$  corresponding to low or negligible effective memory, the particles have effectively zero velocity. When  $M$  increases beyond a critical threshold (solved for numerically and labelled in cyan), the particle experiences enough gradient force to move ballistically away from its past locations. As  $M$  increases further, the velocity of the particle rapidly converges to the theoretical value of  $V = 6$  satisfying Eq.[original velocity equation] when  $\mu = 0.011$  and  $\nu = 998.5$ , and  $\phi = 1$ .

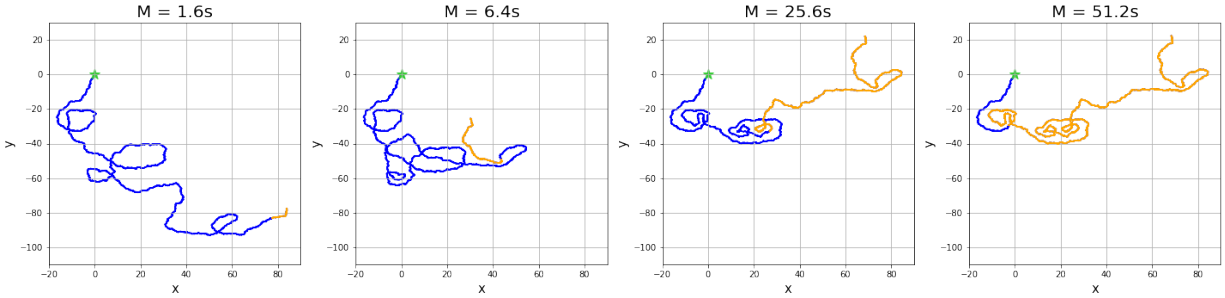


Figure 4.15: As the memory integral window size  $M$  increases, new path dynamics emerge. For a particle starting at  $(0, 0)$  (green star), each panel shows a new path with progressively longer effective memory (larger  $M$ ). Gold path sections show the length of path used corresponding to  $[t - M, t]$  for  $t = 60$  (i.e, for the particle's next step at  $t = 60 + dt$ , how much past history the particle has access to). At low  $M$  the paths still exhibit curvature, but at higher values of  $M$  the paths begin to self-trap. It is interesting to note that for effective memory window  $M = 26.6s$  and  $M = 51.2s$ , the path dynamics are visually identical. For each path realization, we use the same translational white noise at each timestep to ensure that the only variation between paths occurs as a result of the changing integrand in Eq. 4.2 controlled by  $M$ .

same  $\mu$  and  $\nu$  and  $V$ . (We consider the ensemble because even by re-using the same random noise as we did in Fig. 4.15, we produce totally different paths when we change  $M$ , therefore at most we can compare how the average behavior of the ensemble changes.) For the straightness data of each ensemble  $\{S_n\}_M$

we compute the ensemble average mutual information decay curve and extract the decay timescale  $\alpha$ . We track how the average mutual information decay timescale  $\alpha$  changes with the memory tuning parameter  $M$  in Fig. 4.16. As we might expect, looking back further in time (increasing the effective memory  $M$ ) correspondingly increases  $\alpha$ , which is the strength of the nonlinear correlations between past and present curvatures of the ensemble. Therefore, we conclude that changes in  $M$  directly affect the memory and this change corresponds to average changes in the MI decay timescale  $\alpha$ .

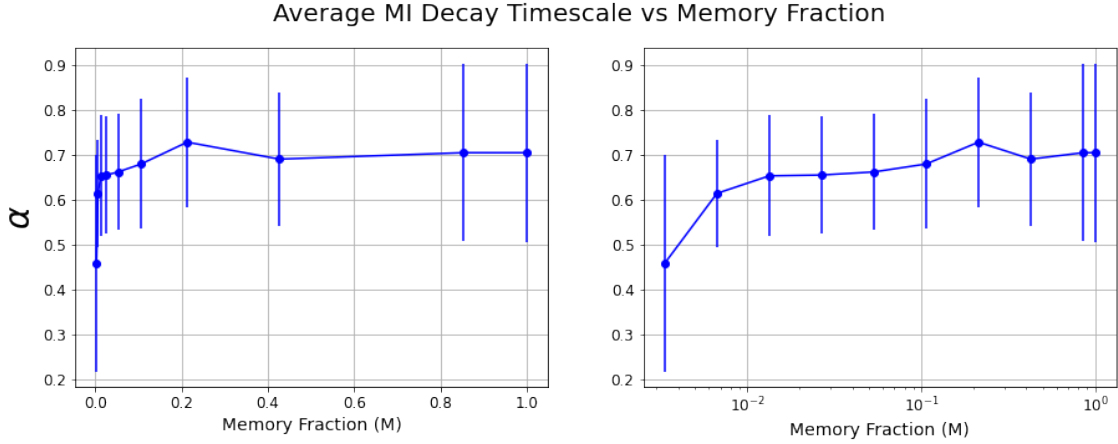


Figure 4.16: The mutual information decay curve of each path is computed and then fitted for the decay exponent,  $\alpha$  as a function of the memory truncation length,  $M$ . The average decay exponent  $\alpha$  increases as the effective memory  $M$  increases. Here,  $M$  is recorded as the fractional amount of past history that the particle has access to relative to the path length. (Therefore,  $M = 1$  is the full memory limit corresponding to Eq. 4.1.) Left plot has standard axes, right plot shows the trend on semilog axes to expand the values near zero.

## 4.7 Comparison Across Models

We showed previously that the mutual information of the path curvature in the ballistic-timescale regime can be used to understand the self-avoidant memory response of active particles within the same model class. A natural question that arises from this finding is whether or not the mutual information of path curvature can distinguish particles from different model classes. To this end, we will compare the mutual information of the curvature of our self-avoidant particles against two other models which can be tuned to have varying levels of path curvature in the ballistic-timescale regime:

1. *Self-avoidant particles*: no explicit curvature source within the model equations,

2. *Active Brownian particles*: no explicit source of curvature, but the relative straightness of paths may be controlled by tuning  $\tau$  (choosing large  $\tau$  creates straighter paths),
3. *Active chiral particles*: explicit curvature controlled by  $\omega$ .

We have discussed the equations of the active Brownian model at length in chapter 2. The active chiral model can be viewed as an extension of the active Brownian model with an added deterministic parameter,  $\omega$ , which biases the random reorientations to trace out circular-looking paths with the following equations:

$$dX = V \cos(\theta(t))dt + \sqrt{\epsilon}dW_x \quad (4.5a)$$

$$dY = V \sin(\theta(t))dt + \sqrt{\epsilon}dW_y \quad (4.5b)$$

$$d\theta = \omega + \frac{1}{\sqrt{\tau}}dW_\theta. \quad (4.5c)$$

(Note that choosing  $\omega = 0$  will recover the active Brownian model.)

We show the path information, straightness index time series, and resulting fitted average model mutual information decay curve in Fig. 4.17. For the self-avoidant sample path (blue), we have selected a path which traps itself twice with our usual parameters  $\mu = 0.01$  and  $V = 6$ . For the active Brownian path (green), we choose a comparable velocity to the self avoidant paths ( $V = 6$ ), but choose  $\tau = 1/5$  to force fast reorientations which will decrease the straightness of the path substantially. The active Brownian path data shows that the active Brownian particle covers much less space than the self-avoidant particle, due to these frequent changes in direction. The effect of frequent directional changes can be seen in the straightness index time series of the active Brownian path as well- the straightness time series oscillates mostly between 0 and 0.6 in comparison to the self-avoidant path which oscillates between 0 and 1. In contrast, the chiral particle has a consistent spiral path structure which generates a straightness time series that is the lowest at all times and which never is higher than 0.5.

The bottom panel of Fig. 4.17 shows ensemble average time-delayed mutual decay curve of the straightness data for  $N = 24$  paths for each model. These curves are fitted to a decaying exponential  $e^{-\frac{t}{\alpha}}$ . The self-avoidant model has both larger and longer lasting nonlinear correlations over all time delays  $T$ , even compared to a model with “many small curves” (active Brownian, for the parameters chosen) and also compared to a model with many large curves (active chiral, for the parameters chosen). Thus, we find that the

mutual information can distinguish “high curvature” at large and small scales (active Brownian and active chiral) from curvature which arises from self-interactions.

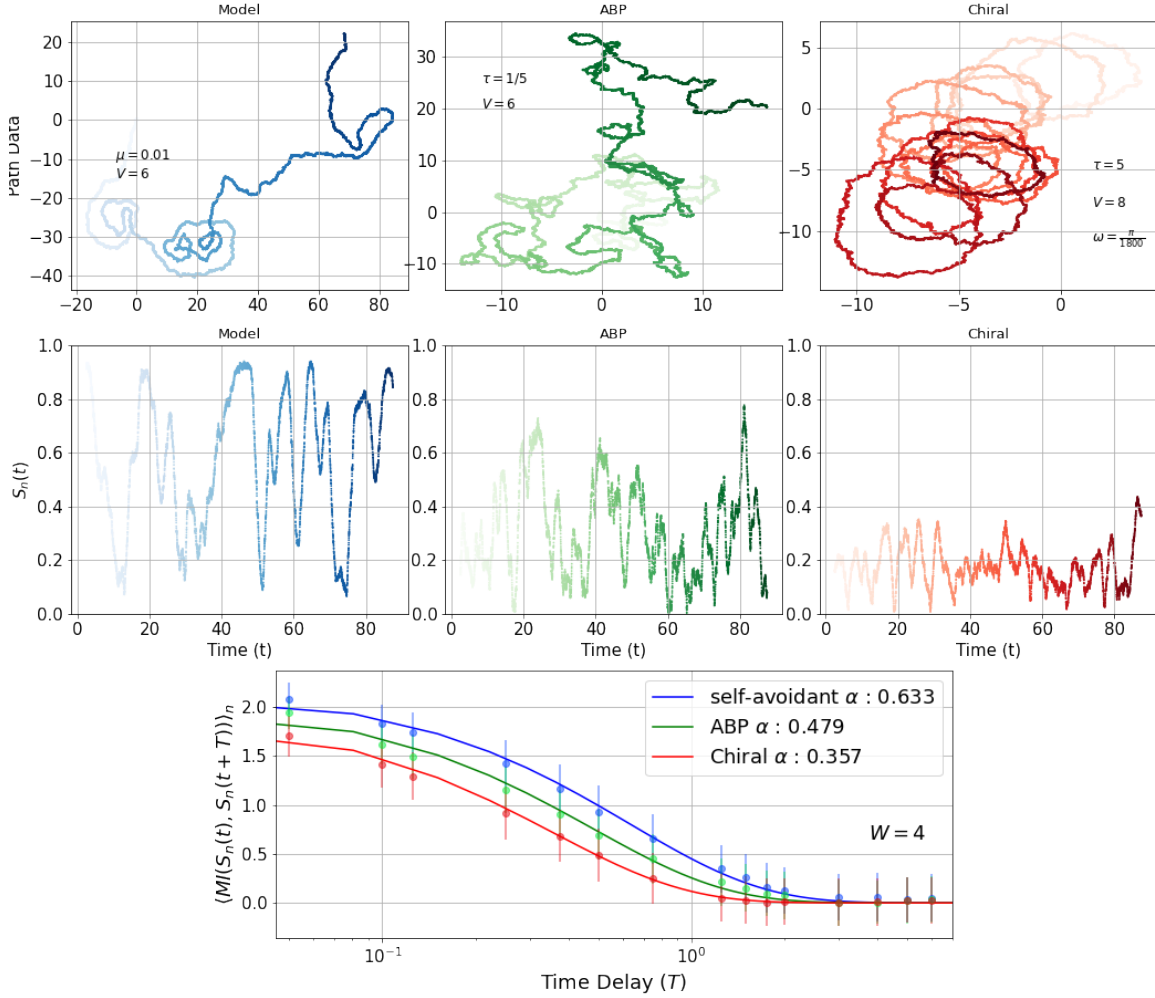


Figure 4.17: Comparison of mutual information of curvature between three models with different curvature sources. Overall time delays  $T$ , the average mutual information curve of the self-avoidant model takes on higher values and a longer decay timescale when compared to the active Brownian model with fast reorientations and compared to the active chiral model with long slow turns.

## 4.8 Conclusions

We have shown that the self-avoidant memory response of theoretical self-avoidant particles is not well explained by conventional statistical methods for analyzing path data. Additionally, conventional active particle models such as the active Brownian model are unable to reproduce the characteristic path features that we observe in self-avoidant particle paths, specifically, self-trapping. In general, *average reorientation*

time can be estimated by fitting the velocity autocorrelation function, but the observed self-traps are created by *consecutive reorientations in the same direction* which produce path segments with high curvature. New tools for understanding this were required.

Since existing methods do not suffice to adequately characterize the path dynamics which we observe, we reimagine our path data as a straightness index which we find has temporally decaying nonlinear correlations with itself. These correlations are revealed by computing the time-delayed self mutual information using an empirical estimation scheme which must be adapted to accommodate dynamical correlations within the time-series straightness data. We find that this method convincingly distinguishes different levels of self-interaction between individual paths that are generated with the same model parameters: paths which exhibit higher levels of self-interaction have higher and longer lasting nonlinear correlations in straightness data.

For individual paths with different visual features suggesting different levels of self-avoidant memory response, we showed that these differences can be distinguished using the mutual information of path curvature. We continued by introducing an effective memory timescale  $M$  which functioned to control the amount of past history “seen” by the particle at time  $t$ , which amounted to all past locations in the interval  $[t - M, t]$ . By increasing the size of  $M$  and introducing longer memory into the path data we found that the mutual information decay timescale  $\alpha$  also increased, and therefore longer memory times induced longer nonlinear correlations in the path curvature, on average. This provided additional evidence that the mutual information decay was able to distinguish different levels of self-avoidant memory.

In our final section, we also investigated differences in the mutual information of path curvature between different models. Specifically, we compared our self-avoidant model to the active Brownian model with fast reorientations using small  $\tau$  to produce many small areas of curvature and to the active chiral model with slow turning rate  $\omega$  to produce large and consistent spirals. In this comparison to two different manifestations of curvature (“small” and “large”), we found again that the self-avoidant particles had both larger and longer lasting nonlinear correlations in the curvature.

We conclude by restating our original hypothesis, which is that the self-traps created by the self-avoidant memory response can be identified by regions of high curvature, and that this curvature is an emergent feature of the memory. Using a KNN mutual information estimator adapted for time series data, we provided evidence supporting our hypothesis by distinguishing both individual paths on the basis of their levels of self-interaction and by distinguishing our model from other models with curvature.

## 4.9 Future Work

In our study so far, we have considered the mutual information of the straightness index  $S_n(t)$  with itself at future times,  $S_n(t+T)$ . We argued that the self-trapping is marked by curvature, which can be described as repeated reorientations in the *same* direction, and that this curvature was the more important feature to consider, rather than only changes in the direction of the particles (which reorientation timescale  $\tau$  can quantify). In future work, we may explore increasing the dimension of the variable  $\{Z_i\}$  by including the orientation,  $V_n(t)$ . Instead of considering  $MI(S_n(t); S_n(t+T))$ , we may consider  $MI((S_n(t), V_n(t)); (S_n(t+T), V_n(t+T)))$ .

Other directions include a more detailed exploration of the parameter space of the self-avoidant model to explore whether there exists a quantifiable parameter regime that can produce self-trapping. We also will consider increasing the total number of paths studied for fixed parameter sets to generate better statistics and investigate whether the true distribution of mutual information decay timescales  $\alpha$  can be estimated.

## CHAPTER 5

### Using Trajectory Data to Infer the Character of Social Interactions Between Golden Shiners

#### 5.1 Introduction

It is useful to understand signaling and communication pathways between organisms, but the internal cognitive state of non-human organisms can be difficult if not impossible to ascertain. What little information we may be able to gain about an organism's internal cognitive state may be even more difficult to interpret. To this end, we propose a method for estimating the extent to which one golden shiner's movements are related to another's using nonlinear statistical correlations within their collective movement data. Golden shiners (*Notemigonus crysoleucas*) are gregarious organisms, meaning that they prefer to live in same-species groups (shoals) and are highly sociable. These shoals function to protect juveniles, help defend against predators, and increase foraging capability in the wild. In experimental settings, it has been shown that a minority of individuals may influence the behavior of a shoal [57]. It has also been shown [39] that in golden shiner shoals select "leader" individuals are usually positioned near the front of the group. A study performed by [40] further quantifies that individuals may act as leaders consistently in an experimental setting. Various theories have been posed to identify the morphological or metabolic conditions that are correlated with which fish are leaders.

While some studies have found correlations within the movement data of animals who are observed to engage in leader-follower-like behaviors [6], [51], [56], the literature is left wanting for quantitative evidence that these observed "leader-follower" configurations and the associated correlations are not merely coincidental. Consider runners racing in a 100m dash- the foremost runner, or "leader" is positionally ahead of the rest of the racers, but the runners behind the race leader are all running independently to the finish line. In our experiments, we remove all extraneous environmental factors and assume no relationship between individuals; therefore we assume no correlated motion in any sense. Just as positionally "being in front" does not mean there is a leader-follower relationship, it is also true that correlated motion could be coincidental (i.e., correlation  $\neq$  causation). To surmount the issue of coincidental correlations, sophisticated

information theoretic sophisticated metrics such as transfer entropy have been developed to statistically isolate directional interactions from other sources of correlation. While these metrics theoretically prove that one agent's movement decisions depend upon the motion of another agent, such metrics have two main drawbacks. The first is that they require an assumption that correlated motion implies social interaction, and the second drawback is that these metrics require conditioning on past states of the organism, which is difficult to implement with experimental data. In our approach, we make no assumptions of social interaction and compute both two-point linear and nonlinear correlation (mutual information) metrics to compare our experimental trajectory data to a reduced toy model of with an explicit leader-follower relationship between agents. The structure of the nonlinear correlations that we find using the time delayed mutual information is in good agreement with this explicit leader-follower model. Since our findings are consistent with this leader-follower interpretation, results suggest that leader-follower interactions are the dominant social force that drives the observed coupled motion between golden shiners.



Figure 5.1: Golden shiner in a person's hand, taken by Amy Schrank [46].

## 5.2 Experimental Methods

Our collaborators at the Cognitive Ecology and Behavioral Engineering Laboratory (BEL) of the U.S. Army Corp of Engineer's Engineering Research and Development Center (ERDC) in Vicksburg, Mississippi performed the data collection and experimental management. Outside of experimentation, four hundred juvenile golden shiners were housed in an indoor holding tank with fluorescent lights on a 12h light/dark cycle. Fish were fed once daily *ad libitum*, after experimental individuals had been selected. To ensure a sterile environment, water quality measurements were taken daily, including temperature, dissolved oxygen,

pH, conductivity, and oxidation-reduction potential (ability of water to break down contaminants). Additional weekly tests were performed to measure ammonia and nitrate content as well as carbonate hardness.

Two types of experimental conditions were tested: a control state and an agitated state, which was implemented by enforcing two environmental conditions that have been shown to induce behavioral changes in golden shiners. The first condition was a reduced water depth; the water depth level of the experimental tank was reduced to half of the control depth (from 8cm to 4cm). Changes in water depth have been shown to prompt fish to behave defensively by tightening the shoal formation to reduce individual risk from predators [7], [27]. Golden shiners have been also been shown to prefer shaded regions which may be a strategic move to minimize the ability of predators to see them [78], [5]. Accordingly, we increased the overhead ambient light from 200 lux in the control setting to 230 lux in the agitated setting.

Experimental pairs were selected randomly and taken to the experimental tank in a separate temperature controlled room. An aerial view of the annular tank setup is shown in Fig. 5.2. The inner diameter of the tank measured 23.16 cm and the outer diameter measured 125.1 cm. Once (unfed) experimental pairs were transferred to the tank, the fish were left undisturbed for 10 minutes to acclimate to the new environment. After the acclimation period, video data was recorded using a Basler Boost Monochrome high resolution camera for a period of 30 minutes and at a frame rate of 40 frames per second at a resolution of 2912 x 2750 pixels. The still images (frames) of the video data were converted into a continuous video format using the ffmpeg library. Four replicate experiments were performed in the control setting, and five experiments were performed in the agitated setting. At the conclusion of each 30 minute experiment, fish weights and lengths were measured before euthanization.

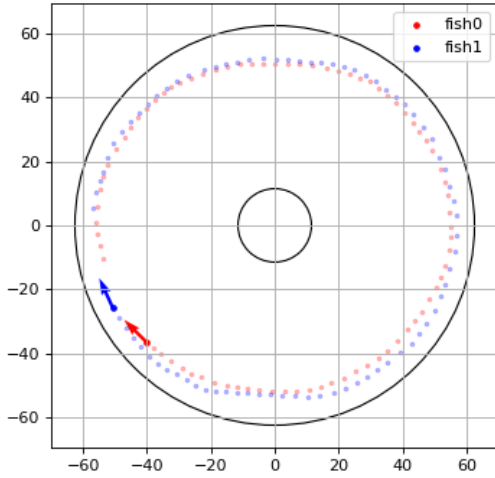
The collected video data shows four distinct behaviors:

1. **Smooth laps**, in which both fish swim around the tank in generally smooth circles. Usually there is a fairly obvious positional leader. Often characterized by high velocity.
2. **Erratic laps**, in which fish swim around the tank in irregular circles at varying speeds. May be punctuated by short periods of localized meandering or stillness.
3. **Localized meandering**, in which fish remain close to each other and confined to a small area of tank for a distinct period of time. Often one fish swims small circles around another.
4. **Stillness**, in which both fish are still.

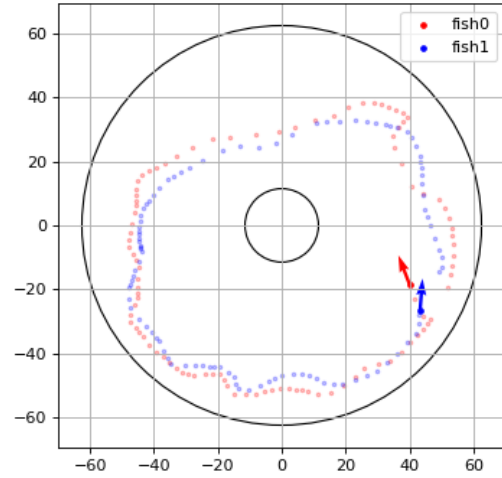
Trajectory segments corresponding to these behaviors are shown in Fig. 5.3. Note that different behaviors in each panel are captured over varying lengths of time. Although the four main categorized behaviors are distinct, the video data shows that both fish spend a majority of each experiment cycling through the different modes together. For example, the fish may both swim smooth laps before both transitioning to localized meandering. This coupling suggests a robust and consistent social interaction between fish. Furthermore, we identify consistent asymmetries in the movement data of the three motion-driven states: during the smooth lap and erratic lap state, one fish appears to be positioned ahead of the other consistently, and during localized meandering, one fish appears to swim around the other, mostly still fish. These asymmetries indicate that perhaps there is a substantial leader-follower element within the social interactions we observe in our experimental data. We suppose that the smooth lap leader-follower state is the state with the highest level of interaction between fish, followed by erratic laps, localized meandering, and stillness. We aim to quantitatively measure this leader-follower dynamic using mutual information derived only from the trajectory data, with no other assumptions.



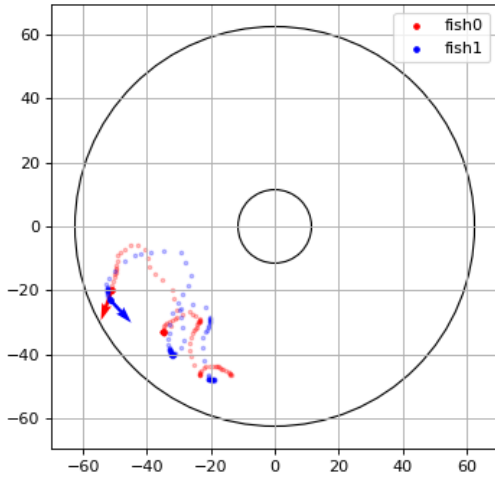
Figure 5.2: Still aerial image of the annular experimental tank, with two golden shiners present.



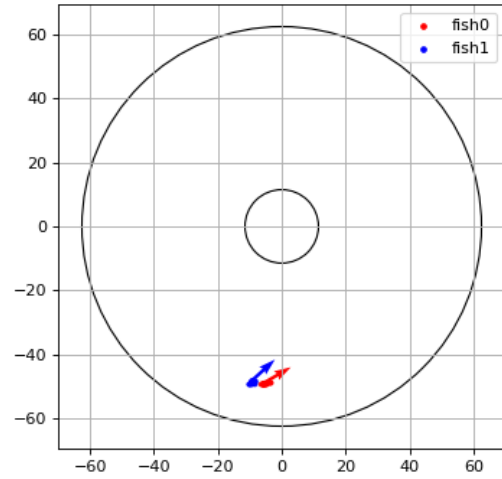
(a) Smooth laps



(b) Erratic laps



(c) Localized meandering



(d) Stillness

Figure 5.3: (a) Smooth laps, over a time period of 10 s during experiment AS-2, (b) Erratic laps, over a time period of 25 s during experiment AS-2, (c) Localized interaction, over a time period of 30 s during experiment CS-5, (d) Stillness, over a time period of 20 s during experiment CS-5.

### 5.3 Data Preparation and Examination

The captured video data was transcribed using the image tracking software TRex. TRex is an open source platform which leverages computer vision and machine learning to identify and track moving entities (Walter and Couzin, 2021 eLife). With TRex, the posture of each individual fish was transcribed into a position time series featuring a centroid location and head location at each frame for each individual fish. We removed small amounts of data where the image tracking software did not record positional data. We label our agitated experiments as AS-1, AS-2, AS-3, AS-4, and AS-5, and the four control experiments as CS-2, CS-3, CS-4, and CS-5.

After re-centering our data so that the origin of our coordinate system was the tank center, we transformed the raw centroid and head location time series data several ways to calculate variables that better represented the mathematics of our problem. From this centered data the angular position  $\theta_i$  of each fish was computed as the two argument inverse tangent function  $\arctan2$  which computes angle between a position vector and the positive x-axis in radians. (The function  $\arctan2$  returns an unambiguous angle between  $[-\pi, \pi]$  by accounting for the quadrant of the positional vector.) To compute the angular position, the centroid position vector was used, meaning that the angular positions  $\theta_i$  represent the angular position of the centroid of each fish relative to the positive x-axis. The individual headings  $\psi_i$  were computed as the angle between the positive x-axis and the positional vector from the centroid to the head of each fish also using  $\arctan2$ . We found camera errors in which the centroid and head locations were mismatched, resulting in consecutive headings that varied by nearly  $180^\circ$ . To address this, we filtered out consecutive headings which varied by more than  $|\pi - 0.6|$ .

We also computed the velocities of the  $\theta$  and  $\psi$  variables by taking the difference of consecutive elements in radians, divided by the time difference:  $\frac{d\theta_i}{dt}$  and  $\frac{d\psi_i}{dt}$ . To ensure that we computed the minimum angle between consecutive headings and angular positions, we solved for the angular difference  $w$  using the dot product formulation between vectors  $\vec{u}$  and  $\vec{v}$ , which is  $w = \arccos \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| |\vec{v}|}$ . (In this way, we assume the smoothest time evolution of the  $\theta$  and  $\psi$  variables.)

Additionally, a relative heading angle, which we call the alignment angle  $A_{ij}$ , was constructed to quantify approximately how much fish  $i$  is oriented in the direction of the fish  $j$ . To do so, we compute the centroid connecting vector, which is the vector pointing from the centroid of fish  $i$  to the centroid of fish  $j$ , and call this vector  $C_{ij}$ . Then,  $A_{ij}$  is computed to be the angular difference between  $C_{ij}$  and  $\psi_i$ . Similarly,  $A_{ji}$  is

the difference between  $C_{ji}$  and  $\psi_j$ . By construction, completely aligned (collinear) fish in a leader-follower formation would have alignment angles of  $\pi$  for the leader and 0 for the follower. We demonstrate this configuration in the schematic of Fig. 5.4.

We find that the alignment angle evolution throughout the experiment is useful for classifying the general dynamics of each experiment. The data from agitated experiment AS-2 in Fig. 5.4 shows that between the time period of approximately  $t \approx 1100$  and  $t \approx 1350$  there is a strong polarization of the alignment angles with  $A_{10}$  trending very close to  $\pi$  and  $A_{01}$  trending close to 0. Video data confirms that during this time period, the fish are swimming smooth laps at an increased angular velocity in a leader-follower formation where fish1 is the leader. The “smoothness” of these laps can be verified by considering the low variance of  $A_{01}$  and  $A_{10}$  during this time period. After  $t \approx 1350$ , we see an abrupt switch in leadership, and then a continued leader-follower state (although slightly less smooth) until about  $t \approx 1500$ , where there is another switch in leadership. In Fig. 5.6, we look at data from agitated experiment AS-3 and observe that the alignment angles are mostly polarized throughout the entire experiment, with fish1 leading the majority of the time. Overall, however, the velocities are lower when compared to the previously discussed segments of experiment AS-2. This is reflected in the higher apparent noisiness of the data in Fig. 5.6 as compared to the leader-follower segments shown in Fig. 5.5.

Although the conditions of the agitated experiments were similar, the fish behaviors were not. Fig. 5.7 shows data from agitated experiment AS-1, and we observe very little polarization and generally lower velocities overall. Video data confirms that the fish spend significant time swimming in erratic laps, which is reflected by the noisiness of the alignment data and weak polarization. Finally, in Fig. 5.8 we show data from a control experiment during which there was very little motion. Video data confirms that the fish swim no laps with each other and are either still or locally meandering.

## 5.4 Mutual Information Setup and Variable Selection

In the original presentation of the algorithm, the k-nearest neighbors MI estimation method was presented for use with a static point cloud [32]. For position time series data, we proposed a modification which considers the dynamical correlations that are present between points that are near in time. We discuss this solution in Chapter 2: a modified sampling scheme where consecutive time samples at  $t_i$  and  $t_{i+1}$  are separated by a window  $W$ , where we suggest that  $W$  should be selected to be the minimum window size

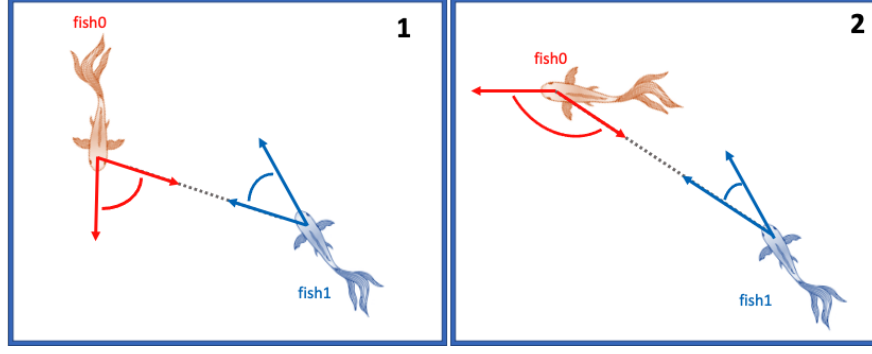


Figure 5.4: In panel 1, fish 0 and fish 1 are facing each other and have similar alignment angles. In panel 2, fish 0 leads fish 1 and fish 0 has an angle close to  $\pi$  where fish 1 has an angle close to 0.

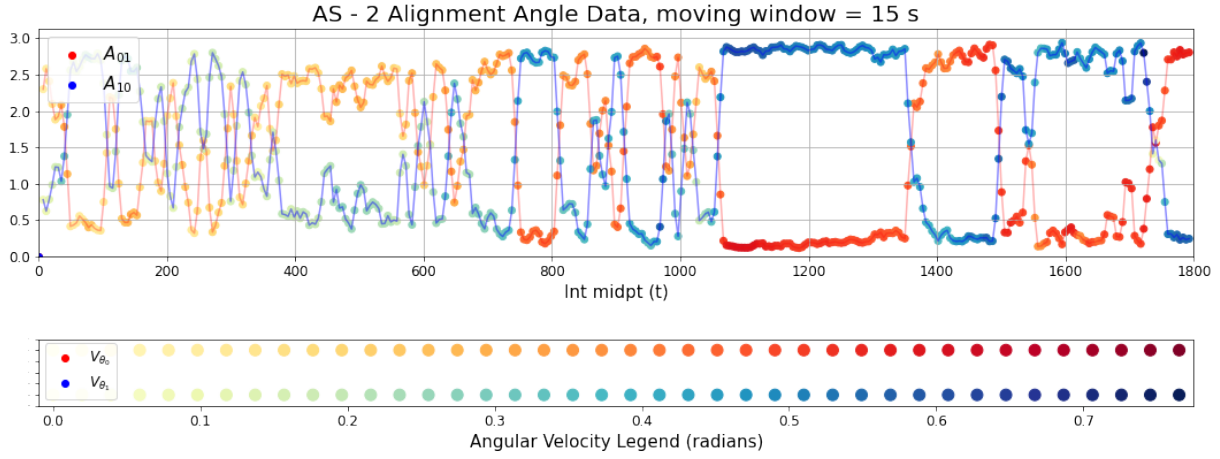


Figure 5.5: 15 sec moving window average of the alignment angle of both fish throughout the experimental time frame of agitated experiment AS-2. Highly polarized time periods, such as the period between  $t \approx 1100$  and  $t \approx 1350$  correspond to a strong leader-follower dynamic and tend to coincide with higher velocities, which is confirmed by video data.

beyond which MI decay curves decay to zero. As before, the time delay  $T$  which captures the time separation between  $X(t_i)$  and  $Y(t_i + T)$  is not constrained by  $W$ . Recall that we computed the *time delayed self mutual information* for individual swimming droplets in Chapter 4, but in this chapter we compute the *time delayed mutual information between fish*. Additionally, we observed in Chapter 4 that our method may only work for confined systems in which the configurational distribution becomes stationary at long times. This appears to be satisfied in our experimental setup since the experimental domain (annular tank) is spatially confined and the experimental data is long enough such that the fish explore the configuration space of angular position well at long times.

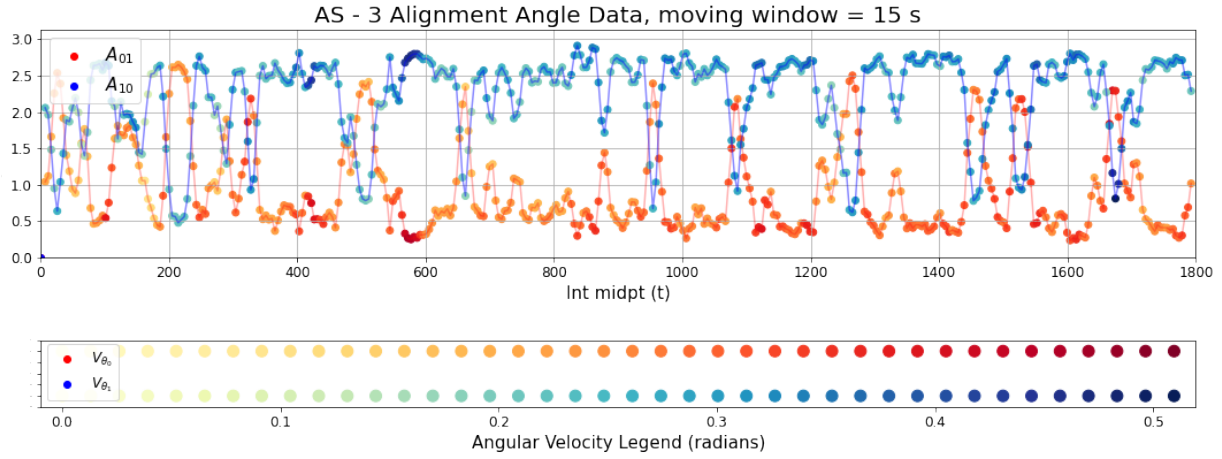


Figure 5.6: 15 sec moving window average of the alignment angle of both fish throughout the experimental time frame of agitated experiment AS-3. The experiment is highly polarized from start to finish, with fish1 dominating the leader-follower laps.

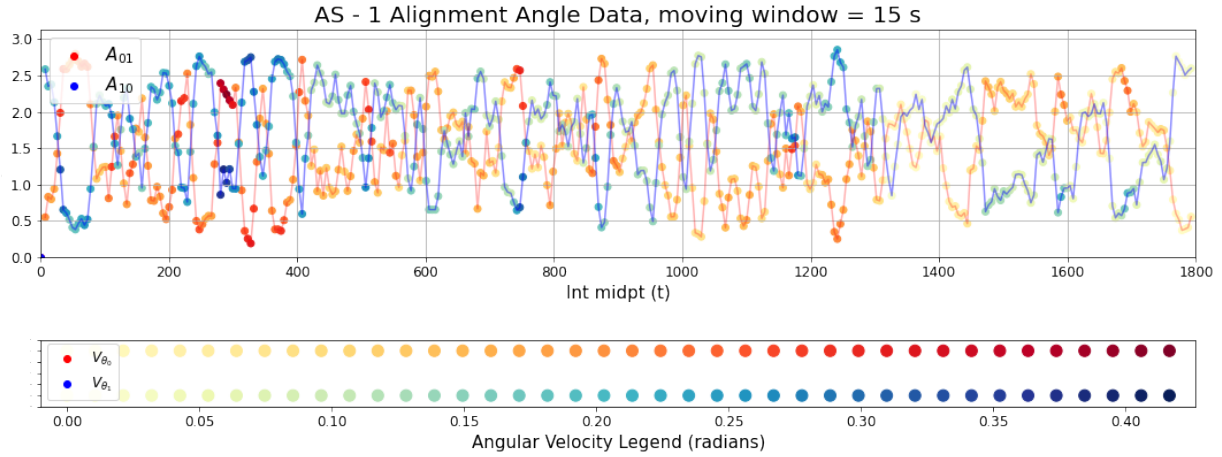


Figure 5.7: Moving window average of the alignment angle of both fish during agitated experiment AS-1. In this experiment, the video data shows few instances of smooth leader-follower laps, which can further be seen in the alignment angle data which is both noisy and not very polarized.

When computing the mutual information of various subsets of our prepared data, it was necessary to pick appropriate random variables to analyze. Early investigations of variables including the rectangular positions, radial position, angular position, heading, and associated velocities yielded insignificant values for non-angular variables and the velocity data. Fig. 5.9 shows the mutual information of each variable pair amongst angular positions, heading, and alignment angles, or  $\theta_i$ ,  $\psi_i$ , and  $A_{ij}$ , respectively. Each reported mutual information is the average of 48 repetitions using the maximum number of samples available with the given

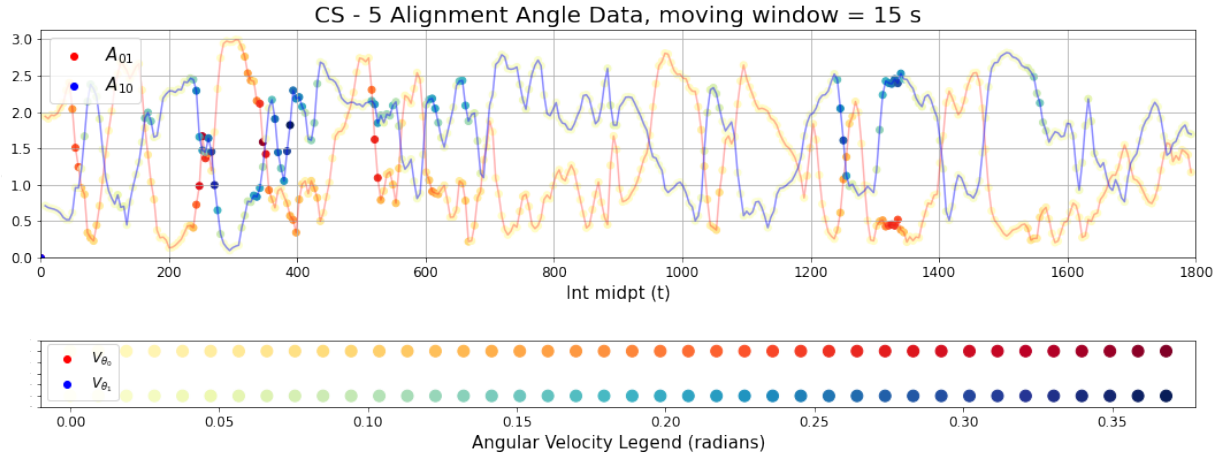


Figure 5.8: Moving window average of the alignment angle of both fish throughout the experimental time frame of control experiment CS-5. In this experiment, the video data shows very reduced motion overall, demonstrated by the generally low angular movement velocity.

window  $W$  and time delay  $T$ . We see that same-variable, same-fish pairs, shown on the diagonals, give the highest mutual information, which is sensible since a variable should be correlated with itself at later times if it evolves as a continuous stochastic process. However, these pairs  $MI(\theta_i(t); \theta_i(t+T))$ ,  $MI(\psi_i(t); \psi_i(t+T))$ , and  $MI(A_{ij}(t); A_{ij}(t+T))$  do not convey information about relationships *between fish*, so we discount these in our analysis. Overall, the same-variable mixed-fish pairs  $(\theta_i; \theta_j)$  and  $(\psi_i; \psi_j)$  give the next highest information over all time lags  $T$  for fixed window size  $W$ . Interestingly, the alignment angle mutual informations were lower than the heading angle and angular position variables in general, despite being very useful in summarizing the behaviors in each experiment as shown in Fig. 5.5, Fig. 5.6, Fig. 5.7, and Fig. 5.8. We also found that mixed variable pairings, such as  $MI(\theta_i(t); \psi_j(t+T))$  could yield substantial values, but chose to focus only on same-variable, mixed-fish pairings for interpretability. In particular, we focus the rest of the analysis on  $MI(\theta_i(t); \theta_j(t+T))$ .

## 5.5 Reduced Model

We restate our hypothesis that the primary driver of social interactions between golden shiners has a substantial leader-follower component. To test this hypothesis explicitly, we propose a toy model to represent the evolution of the angular position  $\theta$  of a noisy leader-follower pairing. Agents in our model will evolve in a consistent leader-follower configuration and there will be no other interaction rules; therefore, agreement between our experimental findings and this toy model will provide strong evidence supporting our hypothesis.

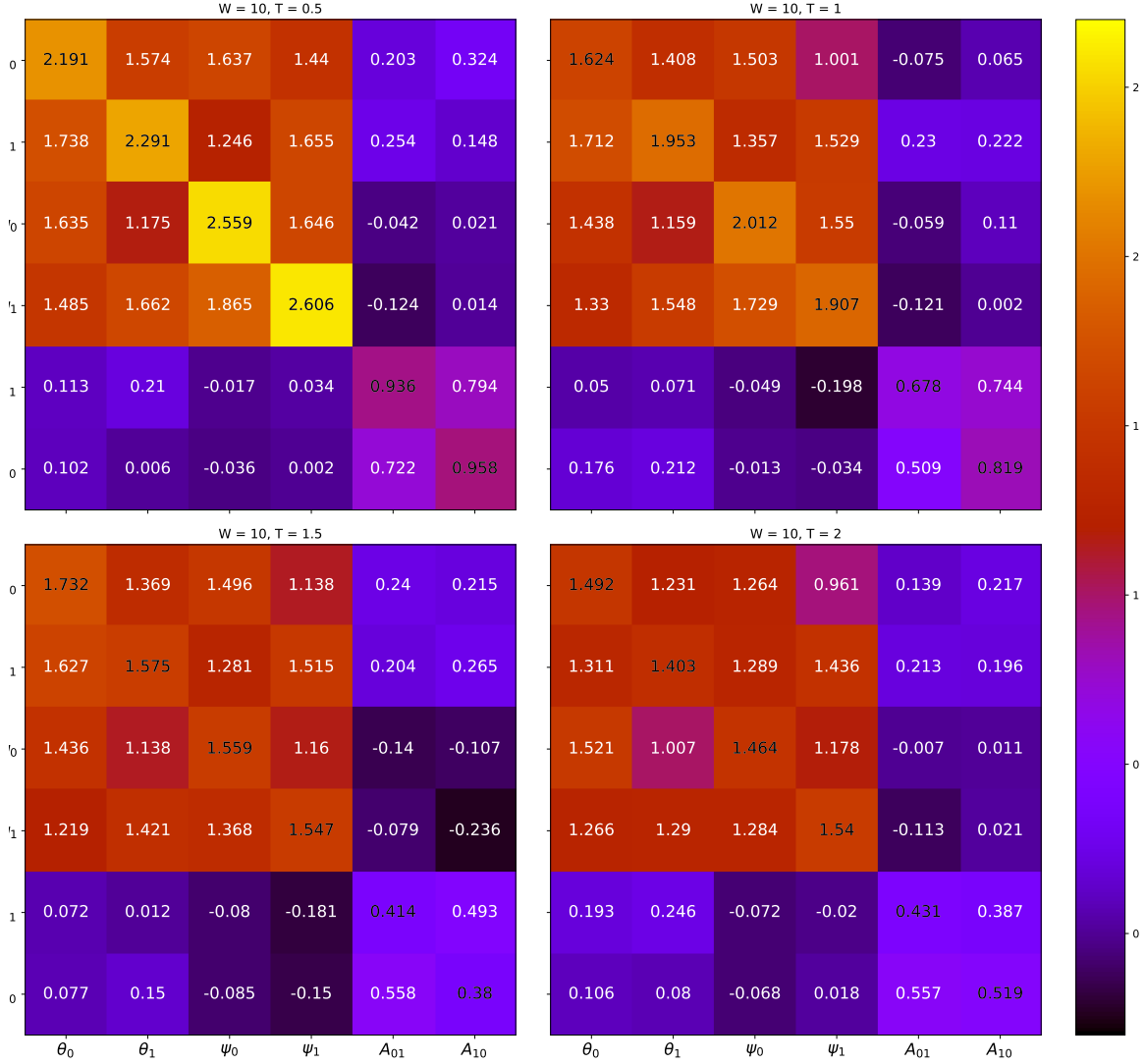


Figure 5.9: Mutual information of selected angular variable pairings of angular position, heading, and alignment angle from experiment AS-2 with a window size of  $W = 10$ . Each matrix represents all mutual information pairs  $MI(X(t); Y(t + T))$  for a specific time delay  $T$ . We explore the asymmetries in the off-diagonals and their implications about fish behavior throughout the rest of this chapter.

We evolve the angular velocity ( $d\theta_L(t)$ ) of the leader using a drift diffusion process, with the following rule to update the angular position  $\theta_L[t]$  at each time  $t$ :

$$\theta_L[t] = \theta_L[t - 1] + V\Delta t + \Delta W_L[t] \quad (5.1)$$

where  $W_t$  is a Wiener process and  $\Delta W_L[t] \sim N(0, \sigma^2 \Delta t)$ . To enforce a leader-follower configuration, we evolve the angular position of the follower by placing it at the leader's position at time  $t - T^*$  and adding

noise:

$$\theta_F[t] = \theta_L[t - T^*] + \Delta W_F[t], \quad T^* > 0. \quad (5.2)$$

The white noise of our follower is modeled by a Wiener process  $W_t$  with  $\Delta W_F[t] \sim N(0, \sigma^2 \Delta t)$  as well. Here,  $T^*$  is the interaction timescale between the follower and the leader; it takes exactly  $T^*$  seconds for the leader to observe the leader's position and react by updating its own position to that same location (plus noise). We stochastically switch the roles of leader and follower using a rate parameter,  $\alpha$ , which is the probability of a leadership switch at each time step. Therefore, the evolution of the angular position  $\theta_i$  follows the dynamics of either  $\theta_L$  or  $\theta_F$  with transitions between the two states governed by the rate parameter  $\alpha$ . We use this model to understand what the mutual information structure of a leader-follower pair *should* look like with signaling timescale  $T^*$ , and absent any other behaviors. Our parameter choices for this model are derived from the experimental data of AS-2. We choose the model velocity to be the average of the distribution of the velocities of both fish0 and fish1, which is  $V = \langle \{V_{\theta_0} \cup V_{\theta_1}\} \rangle$ , where  $V_{\theta_0}$  and  $V_{\theta_1}$  are the sets of all angular velocities of fish0 and fish1 throughout the experiment, respectively. Similarly, we choose our noise strength to be standard deviation of this empirical velocity distribution, which is  $\sigma = SD\{V_{\theta_0} \cup V_{\theta_1}\}$ . For trajectories evolved for  $N$  time steps, we choose the expected number of switches to be  $15 = \alpha \cdot N$ . We enforce periodic boundary conditions on the interval  $[0, 2\pi]$ . A sample trajectory segment is shown in Fig. 5.10.

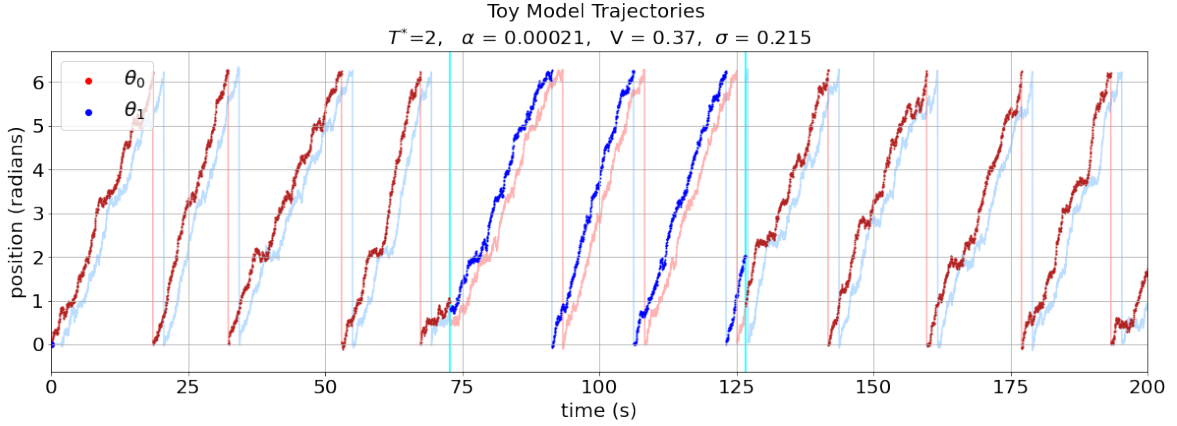


Figure 5.10: Angular position evolution of a toy leader-follower pair swimming in circles for a time period of 200 s, with velocity  $V = 0.37$  and noise strength  $\sigma = 0.215$ . Cyan lines indicate leadership transitions. Leader dark red or blue, follower is light red or blue.

We generated a model trajectory of length  $N = 72000$  and  $\Delta t = 1/40$  for each signaling timescale  $T^*$  value, which was between 0.05 s and 45 s. From these trajectories, we then computed the time-delayed

mutual information between model fish in the same manner as the swimming droplets in Chapter 4. However, unlike before, where we were measuring the *self* time delayed mutual information of a swimming droplet's curvature with its own past history ( $MI(S_n(t_i); S_n(t_i + T))$ ), here we estimate the time delayed mutual information of positional data *between fish*. Specifically, we are interested in whether there are quantifiable asymmetries in two mutual information values:  $MI(\theta_0(t_i); \theta_1(t_i + T))$ , which represents the case of  $\theta_0$  as the leader (since  $\theta_1$  is sampled later in time), and  $MI(\theta_0(t_i + T); \theta_1(t_i))$  which represents  $\theta_1$  as the leader (since  $\theta_0$  is sampled later in time). Since we have assumed that our system is in a steady state, and since the mutual information measures nonlinear correlations between a sample from two presumably stationary distributions of angular positional data, we assume that  $MI(\theta_0(t_i + T); \theta_1(t_i)) = MI(\theta_0(t_i); \theta_1(t_i - T))$ . Therefore, we compute  $MI(\theta_0(t_i); \theta_1(t_i + T))$  across  $T$  values in  $[-45, 45]$ s. To suppress dynamical correlations between consecutive samples of the same variable (such as between  $\theta_0(t_i)$  and  $\theta_0(t_{i+1})$  and between time delayed samples  $\theta_1(t_i + T)$  and  $\theta_1(t_{i+1} + T)$ ), we selected a fixed window size  $W$  which was the average time separation between consecutive samples:  $t_{i+1} - t_i$ . Each data point (y-value) reports the mean and standard error of 48 repetitions where the positional data is sampled with replacement. Separate curves are shown for each fixed value of the separation window,  $W$ .

Fig. 5.11 and Fig. 5.12 show the decay structure of the mutual information computed when sampling the angular position  $\theta_0$  of fish0 first, corresponding to  $MI(\theta_0(t_i); \theta_1(t_i + T))$  on the positive  $T$ -axis, and subsequently on the negative  $T$ -axis we show the mutual information computed when sampling the angular position  $\theta_1$  of fish1 first, or  $MI(\theta_0(t_i); \theta_1(t_i - T))$ . In Fig. 5.11, the true signaling timescale can be seen at  $T^* = 1 = |T|$ s, and in Fig. 5.12 the true signaling timescale is  $T^* = 5 = |T|$ s. In both figures, we note that the maximum mutual information value for both model fish occurs at exactly  $|T| = T^*$  (vertical cyan lines), following a steep decay to zero. This trend is consistent across window sizes. The robust peak at exactly  $T^*$  shows that there is maximal mutual information (nonlinear correlations) between our leader-follower pairing at the true signaling timescale. This provides strong evidence, that, absent any other interactions, the mutual information can recover the true signaling timescale of a leader-follower pairing, which is the location of the maximal mutual information value.

Besides the peak location, the peak heights of the mutual information curves are also interpretable within the leader-follower framework of our model. The peak heights tend to decrease as  $T^*$  increases, which is expected since the model fish will remain positionally closer throughout the experiment for smaller values of  $T^*$  and therefore their motion will be more correlated, yielding higher mutual information values. Recall

Eq. 5.1, in which the random variable  $\theta_L$  is a gaussian random walk with drift, and each realization  $\theta_L(t)$  a sum of gaussian random variables. As we found in Chapter 2, the movement of a gaussian walker adds information to the system at each time step. Therefore,  $\theta_L(t - T^*)$  contains less information than  $\theta_L(t)$  and this difference increases as  $T^*$  increases. Since  $\theta_F(t)$  depends explicitly on  $\theta_L(t - T^*)$  (see Eq. 5.2), the information in  $\theta_F(t)$  is less than  $\theta_L(t)$ , and this discrepancy *also* increases as  $T^*$  increases. Thus, the peak heights, which report  $MI(\theta_0(t); \theta_1(t + T^*))$  and  $MI(\theta_0(t); \theta_1(t - T^*))$ , should decrease as  $T^*$  increases. For a given experiment with fixed  $T^*$ , we also find that the relative peak heights reflect which fish led more often. In Fig. 5.11, fish0 leads approximately 60 percent of the time, which is reflected in a higher mutual information value of about 2.2 bits at  $T = T^* = 1$  vs 1.75 nats at  $T = -T^* = -1$ . In Fig. 5.12, the leadership proportion is more evenly split between fish0 and fish1 (approximately 50 percent of the time for each), which results in very similar peak heights near approximately 1.5 nats.

Finally, we are able to suggest an interpretation for the decay structure of our mutual information curves as a function of  $T^*$ . As the time delay  $|T|$  increases away from  $T^*$  in both directions, the mutual information decreases rapidly (possibly exponentially). We observe that for smaller values of  $T^*$ , the decay appears to take longer to reach zero which reflects longer lasting nonlinear correlations (although higher values of  $T^*$  are decaying from a higher maximum value). This is again expected since smaller values of  $T^*$  indicate that our model fish are positionally closer than larger values of  $T^*$  over all times during the experiment. Therefore, for  $T_a^* \leq T_b^*$  we expect that the true values of both  $MI(\theta_0(t); \theta_1(t + T))|_{T_a^*} \geq MI(\theta_0(t); \theta_1(t + T))|_{T_b^*}$  and  $MI(\theta_0(t + T); \theta_1(t))|_{T_a^*} \geq MI(\theta_0(t + T); \theta_1(t))|_{T_b^*}$ . We show evidence supporting our claim in Fig. 5.16, which shows how the average maximum mutual information value decays noisily as a function of  $T^*$ .

To contextualize our mutual information findings, we also consider the temporal structure of the two point linear correlations in Fig. 5.13 and Fig. 5.14, which are plotted in the same manner/axes as the previous mutual informations. The correlations show decaying oscillations, which we can fit to the form  $e^{-t/\tau} \cdot \cos(\omega t)$ . Fig. 5.15 shows that the oscillation frequency  $\omega$  recovers the average lap time of the model fish. However, it is not obvious how the identity of the dominant leader fish might be extracted from this data. In the mutual information of Fig. 5.11 we saw a dramatic difference in peak heights which corresponded to the leadership proportions of each fish, but this difference is not visible in the linear correlations shown in Fig. 5.13. Additionally for larger values of  $T^*$ , (such as  $|T| = T^* = 5$  in Fig. 5.14) we find it difficult to distinguish any “peak” or maximum value at  $|T| = T^* = 5$  from the oscillations. We note that mutual

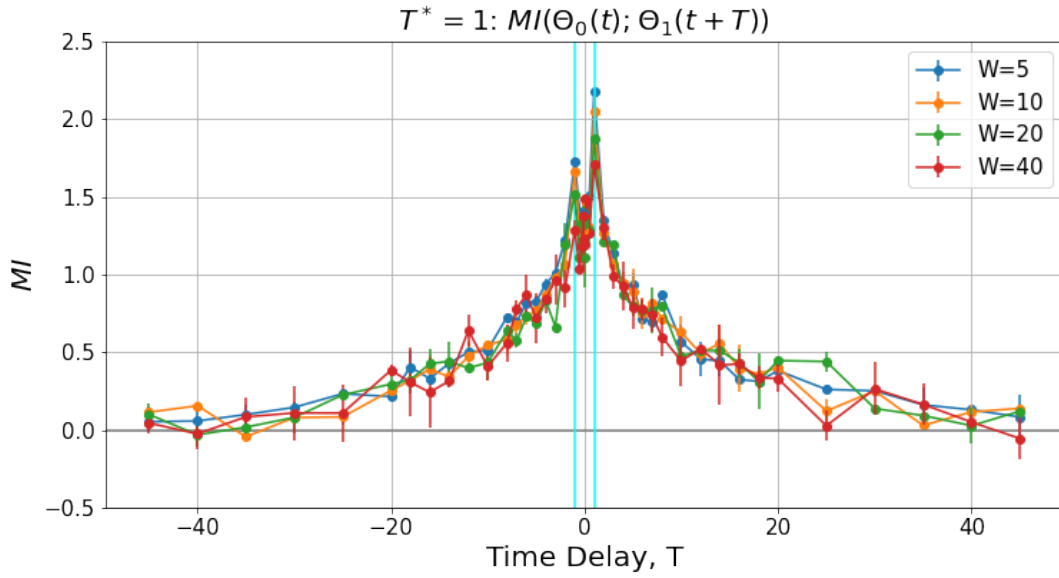


Figure 5.11: When we select  $|T| = T^* = 1$ , we see that the mutual information peaks at  $|T| = T^* = 1$ , then decays slowly to zero. The asymmetry in this peak reflects an asymmetry in the leadership proportions of fish0 and fish1. In this example, fish0 leads 60 percent of the time, and fish1 leads about 40 percent of the time.

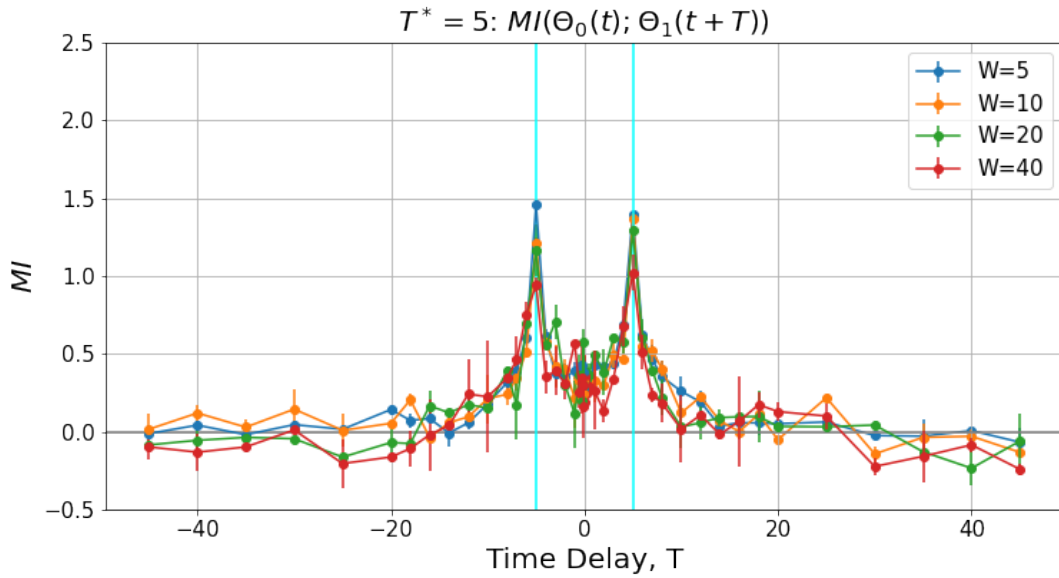


Figure 5.12: Similarly, selection of  $|T| = T^* = 5$  shows a double peak at  $|T| = T^* = 5$ , although the peak heights are lower in both cases. Additionally, the proportion of leadership is evenly split between fish, which can be seen in the matching peak heights.

information decay does not appear to oscillate, and therefore is more likely able to represent the actual nonlinear decay dynamics without being affected by the annular experimental setup.

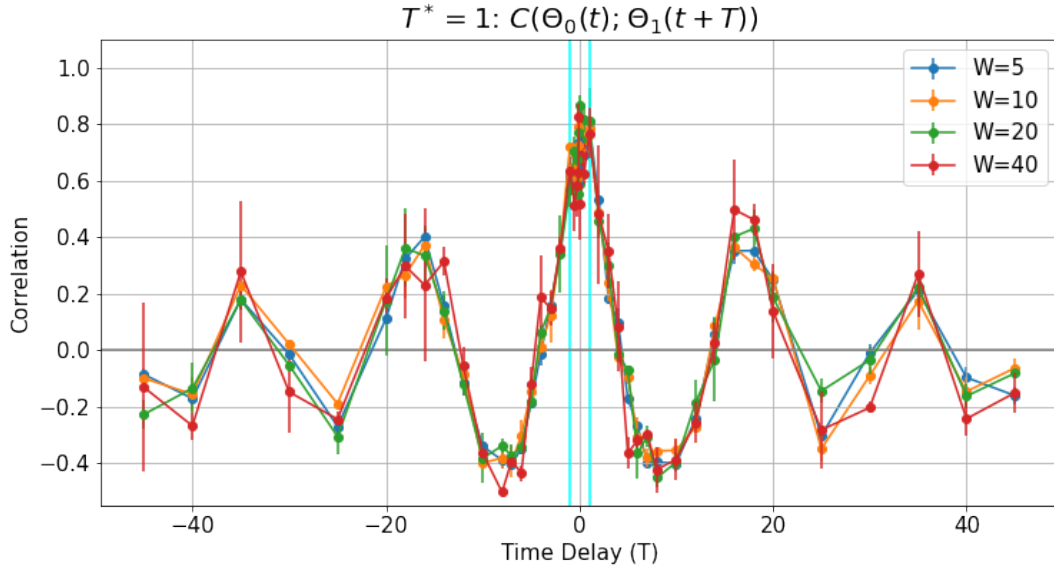


Figure 5.13: When we select  $|T| = T^* = 1$ , we see that the correlation may at  $|T| = T^* = 1$ , followed by slowly decaying oscillations. Reported correlations are the average of 48 repetitions with standard errors.

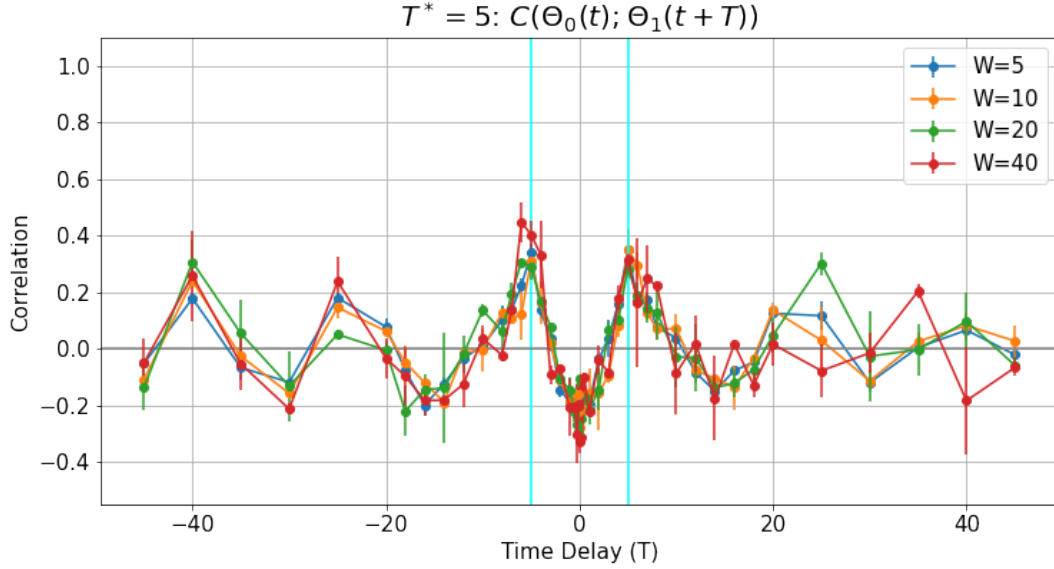


Figure 5.14: Similarly, selection of  $|T| = T^* = 5$  shows what may be a peak at  $|T| = T^* = 5$ , but is difficult to distinguish from the oscillatory peaks.

We have asserted that the peak location in the mutual information curves reveals the true signaling timescale  $T^*$  and that the relative peak heights differences between  $MI(\theta_0(t); \theta_1(t+T))$  reveal which model fish has lead for a greater proportion of the experiment. In Fig. 5.16 and Fig. 5.17, we show the mutual information and correlation values at exactly  $|T| = T^*$ , which amounts to selecting the mutual information

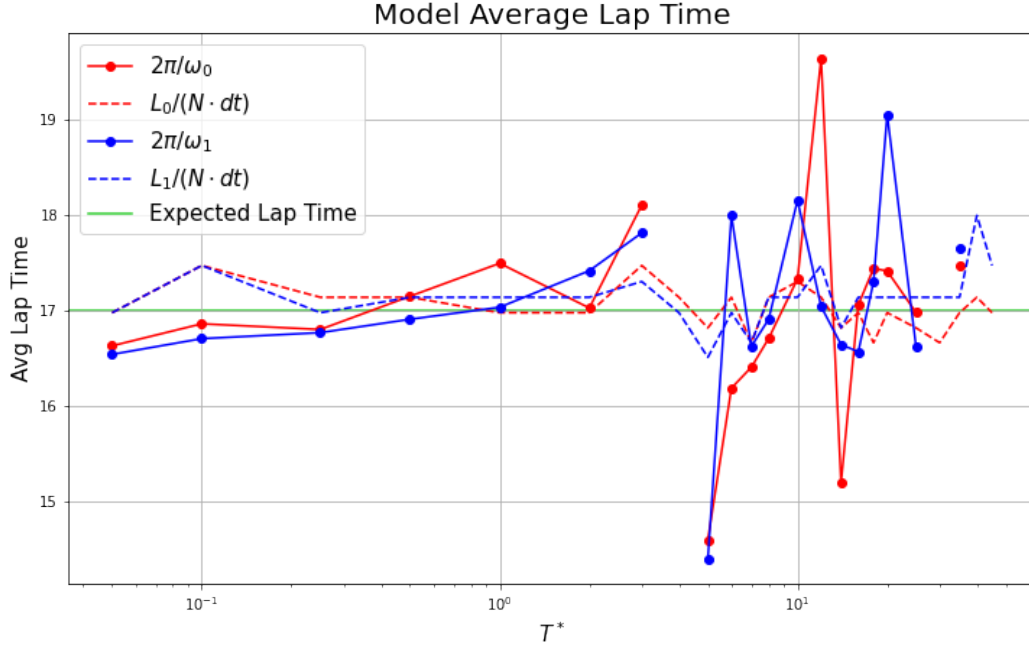


Figure 5.15: Average lap times for model fish as a function of  $T^*$ . The expected lap time (green) is  $\frac{2\pi}{V} = 17$  s. (By construction, the velocity  $V$  in our model is an angular velocity.) Dashed lines represent the average lap time counted from path data where  $L_0$  and  $L_1$  are the number of laps that model fish0 and model fish1 have completed, respectively. Solid lines connect the average value of fitting  $\frac{2\pi}{\omega_i}$  correlation curves corresponding to  $W \in [5, 10, 20, 40]$ . As  $T^*$  increases, the exponential component of the correlation function diminishes and the fits often do not converge, therefore some values are missing.

and correlation coincident with the cyan indicator lines for each experiment with unique  $T^*$ . We see that the mutual information appears to noisily decay exponentially with  $T^*$ , and the curves shift closer toward zero as the sampling window  $W$  increases, which confirms our intuition that samples which are further separated in time (larger  $W$ ) are less correlated. (As  $W$  increases, sample pairs are further separated, and as  $T^*$  increases, the fish are temporally, and thus spatially, more separated.) The linear correlations decay in magnitude, but oscillate rapidly as a function of  $T^*$ , suggesting that the signal given by the correlation curve height at  $T^*$  is possibly just reflecting the annular tank structure. These correlation curves do not reveal the leader-follower dynamic between model fish, and it suggests that the linear correlations are always polluted by the tank structure.

## 5.6 Experimental Data Analysis

We now compute the mutual information over increasing  $T$  and various window sizes  $W$  for our experimental data, in the same manner as we did for our model generated data. Throughout our analysis, we analyze mutual information of the experimental data through the comparative lens of the previously analyzed model, since the model mutual information shows a clearly interpretable leader-follower dynamic. In Fig. 5.18, we assess the mutual information of experiment AS-3 in which the fish swim laps for most of the experiment (as seen in the alignment angles of Fig. 5.6). For window sizes  $W \in [5, 10, 20, 40]$ , the mutual information of experiment AS-3 decays exponentially to zero as the time lag  $T$  increases in either direction, for all values of  $W$ . The largest experimental mutual information values of AS-3 are near 2.25 nats, which is similar to the largest values in Fig. 5.11 corresponding to a true signaling timescale  $T^* = 1\text{s}$  between model fish.

As before, we also compute the linear correlations, shown in Fig. 5.19. We see that the correlations also are high at small values of  $T$ , with oscillations that decay as  $|T|$  increases, which is similar to the model. In the model, the oscillations are much larger and decay more slowly, which we attribute to the fact that the model fish swim *only* smooth laps, whereas the behavior of the experimental fish vary throughout the experiment. We find that the structure of the linear correlations either obscures or cannot report the leader-follower dynamic which the video data revealed, and we continue to explore whether the mutual information and the captured nonlinear correlations are able to better explain the observed system dynamics. In particular, we hope to match the experimental mutual information curve features to those found in our model, since our model could be interpreted in the context of the enforced leader-follower dynamic.

In particular, we hope to find a maximum or peak in our experimental mutual information. In the model, we observed a mutual information peak at the exact signaling timescale,  $T^*$ , however, there is not an obvious peak structure in the experimental mutual information of Fig. 5.18. To investigate further, we increase the density of mutual information calculations at  $|T|$  values near zero; that is, we compute the mutual information as we increase  $|T|$  by a single frame (0.025 seconds) until  $|T| = 2\text{s}$ . We consider experiment AS-3, which shows the most consistent leader-follower state in the alignment data (Fig. 5.6 and video data). This higher granularity of computed mutual information of AS-3 is shown in Fig. 5.20, where there is an obvious peak on the negative  $T$  axis. This visual asymmetry is consistent with the information presented in Fig. 5.6 where we see that fish1 leads for nearly the entire experiment. In the case that the fish swim mostly smooth laps (as vs

other behaviors such as stillness, localized interaction, or possibly even erratic laps), we believe that the trend of the experimental mutual information is consistent with our model findings, indicating that the dominant leader fish will see overall higher mutual information compared to the follower. In our model, we are further able to state the exact timescale of the leader-follower interaction, which is  $T^*$ , and also is the exact location of the mutual information peak. We would like to estimate the value of  $T^*$  from our experimental mutual information data as well. To avoid assuming a peak where there is none (as in the positive  $T$  axis of Fig. 5.20 corresponding to fish0), we will non-parametrically smooth our data using locally estimated scatter plot smoothing, or LOESS [14].

LOESS is a locally weighted regression procedure which uses least squares to minimize the error between scatter plot points and a polynomial fit (therefore, it is non-parametric). LOESS takes in the scatterplot data  $\{x_i, y_i\}$ , as well as a parameter  $f$ , which determines the fraction of data centered about each  $(x_i, y_i)$  which will be used to perform the weighted regression. In Fig. 5.21 we show several LOESS fits of  $MI(\theta_0(t); \theta_1(t + T))$  with  $T < 0$  for AS-3, which we visually know has a strong peak favoring fish1 from the data in Fig. 5.20. We see that for small  $f$ , the fit follows the noisiness of the data, but becomes smoother as  $f$  is increased. For each fit, we can extract the location of the maximum value, which is also reported in the legend. Even for  $f$  as large as  $1/2$  of the data, a max value at  $T = -0.575$ s is still reported, which is substantial evidence that there is a peak in the mutual information on the negative  $T$ -axis for AS-3.

In Fig. C.19, we can see more explicitly how varying  $f$  changes the peak location for experiment AS-3. Each individual plot shows the peak location  $|T|$  as a function of data fraction  $f$  for a given window size  $W$ . For  $W = 5$ , we see a large stratification between the peak locations of fish1 (blue) and fish0 (red) up until approximately  $f = 0.5$ . Beyond  $f = 0.5$ , the location of the peak of fish1 drops to  $|T| = 0.025$ s. The value of  $T^*$  that we estimate is bounded below by the frame rate, which is  $1/40 = 0.025$ s. We interpret this as strong evidence that the true signaling timescale of fish1 in AS-3 is  $T^* \leq 0.575$  s, which is the largest reported value of  $T$  for which a peak is detected as shown in Fig. C.19. Similarly, this evidence shows that the peak value of fish0 is  $T^* \leq 0.025$ . This coincides with Fig. 5.20, where we can see that there appears to be no peak, and the data decays noisily from its maximum value that occurs at the smallest reported timescale  $T = 0.025$ s.

In our model,  $T^*$  represents the time elapsed between leader fish's positional update, and follower fish's positional update *to the leader's exact past position*, plus noise. Agreement between the model and our experimental findings suggests that our estimates for  $T^*$  report the amount of time needed for the signal

(leader positional update) to be transmitted, received, interpreted, and reacted to by the follower, making  $T^*$  the reaction time of the fish. Therefore, experimentally estimated signaling timescales at  $T^* \in [0.025, 0.575]$ s are similar to human reaction times [48].

## 5.7 Experimental Summary Analysis

In Fig. 5.23 and Fig. 5.24, we illustrate the decay structure of the mutual information and linear correlations, respectively, for all experiments over a sampling window size of  $W = 10$  s. For experiments in which the fish are moving for most of the experimental time frame (AS-1, AS-2, AS-3, AS-4, AS-5, CS-3, CS-4), we find that the mutual information follows an apparent exponential decay as  $T$  increases. We note that for experiments where there is little to no movement (CS-2 and CS-5) the curves do not appear to decay within the time delays considered (up until  $|T| = 45$ s). Although those experimental data report long-lasting correlations (slowly decaying mutual information), this is misleading since observation of the system shows little to no coordinated movement of fish. Therefore, although the correlations are high, they are trivial. Since there is little motion, the fish positions are consistently similar, which would make consecutive positions almost perfectly correlated for long time periods.

The correlation decay curves of these experiments decay as well, either in an exponential fashion or in an oscillating fashion. For experiments in which there was mostly lap behavior (AS-2, AS-3, AS-5, CS-3, and CS-4), the correlation function can be fitted to the form  $e^{-t/\tau} \cos(\omega t)$  with lap frequency  $\omega$  where  $\frac{2\pi}{\omega}$  is approximately the average lap speed. These results are tabulated in Table 5.1 where  $L_i$  is the total number of laps swum by fish  $i$  counted throughout the experiment and  $\omega_i$  is the average fitted frequency of correlation curves over  $W \in [5, 10, 20, 40]$ . In control experiments CS-3 and CS-4 the fish swim laps, but at a slower pace than the agitated experiments AS-2, AS-3, and AS-5, which is reflected in lower oscillation frequencies in the control experiments. Like the model correlations, the oscillations of these experimental correlations merely report back the average lap speed of the fish. As in the case with the mutual information, the correlations of CS-2 and CS-5 are trivially high, and do not oscillate since the experimental fish are mostly immobile throughout these experiments. For this reason, we do not include them in Table 5.1.

Experimental Average Lap Times						
Experiment	$\frac{L_0}{N \cdot dt}$	$\frac{2\pi}{\omega_0}$	abs diff	$\frac{L_1}{N \cdot dt}$	$\frac{2\pi}{\omega_1}$	abs diff
AS-2	17.72	15.79	1.93	17.90	15.76	2.14
AS-3	20.81	23.63	2.82	21.06	25.91	4.85
AS-5	17.97	17.17	0.8	17.24	16.27	0.97
CS-3	20.68	21.63	0.95	20.21	21.54	1.33
CS-4	22.51	22.85	0.34	23.71	22.30	1.41

Table 5.1: Experimentally estimated average lap times.

## 5.8 Second Model

Analysis of the experimental data failed to explain why our experimental mutual information curves had softer peaks than those generated by model data. One possible explanation is that the model only evolved the angular position of the fish, without any radial movement, whereas the experimental fish evolved in both the radial and angular coordinates. We propose a modified model in which the leader dynamics are unchanged from the previous iteration, but we increase the follower noise, yielding the following rules. For the leader, as before:

$$\theta_L[t + 1] = \theta_L[t] + V\Delta t + \Delta W_L[t] \quad (5.3)$$

where  $W_t$  is a Wiener process and  $\Delta W_L[t] \sim N(0, \sigma^2 \Delta t)$ . We evolve the angular position of the follower by placing it at the leader's position at time  $t - T^*$  and adding noise which we scale by a factor  $s$ :

$$\theta_F[t + 1] = \theta_L[t - T^*] + s \cdot \Delta W_F[t], \quad T^* > 0 \quad (5.4)$$

where  $W_t$  is a Wiener process and  $\Delta W_F[t] \sim N(0, \sigma^2 \Delta t)$  as well. As before, the angular position of model fish  $\theta_i$  follows the dynamics of either  $\theta_L$  or  $\theta_F$  and transitions between these two states are governed by the rate parameter  $\alpha$ .

We aim to use this modified model to generate mutual information decay curves with “softer” peaks which would be more similar to the experimental data. In Fig. 5.25, we can see that the peaks at  $|T| = T^* = 0.25$  are difficult to see visually when  $s$  reaches only 2. Beyond  $s = 2$  the peak at  $|T| = T^* = 0.25$  is nearly indistinguishable. To successfully obscure the peak for larger values of  $T^*$ , we find that we need larger noise scalings. In Fig. 5.26 and Fig. 5.27 where  $|T| = T^* = 0.5$  and  $|T| = T^* = 1$  respectively, we find that a scaling of at least  $s = 5$  is necessary to make the peaks less visible. While it appears that increasing the

follower noise will make the peaks less visible, it also has the secondary effect of lowering the overall curve height as compared to the experimental data. In general, the experimental data has higher maximum values for very small  $|T|$  compared to the model mutual informations with scaled noise. Therefore, we conclude that further modifications or additions the model would be necessary to better reproduce the trends seen in the experimental data.

## 5.9 Conclusions

We collected experimental trajectory data of pairs of golden shiners in an annular tank under two different experimental conditions (control and agitated). Our goal was to characterize the nature of interactions between the fish. From video data, we found four major recurring behaviors and observed that the fish often cycled through the same behavioral modes together. Positional asymmetries in these behavioral modes (one fish positioned in front of the other) suggested a robust leader-follower dynamic. To characterize this dynamic, we created a toy model that *only* contained leader-follower behavior in the model fish, and found that structure of the time delayed mutual information between model fish recovered the signaling timescale parameter  $T^*$  from the model. This signaling timescale  $T^*$  revealed itself as the location of a sharp peak (and local maximum) in the mutual information curves.

Informed by our model, we estimated the time delayed mutual information between experimental fish positions and found that peaks were difficult to find in the experimental mutual information curves. Using a non-parametric data smoothing technique (LOESS), we were able to recover estimated peak locations in several experiments; in particular experiments where we observed a strong and long lasting leader-follower dynamic in the video data. These estimates for the signaling timescale  $T^*$  were similar to human reaction times.

We found that by increasing the noise of the follower in the model, we were able to produce model generated mutual information curves with softer peaks that were more similar to those of the experimental data. This led us to suggest that the radial motion of the experimental fish, which is not considered in our model, could be causing the experimental mutual information signal to be noisier than the model. Finally, the two point correlation curves of both the model and experimental data revealed little about the dynamics; the only recoverable information from the correlation data was the average lap time, which was found by estimating the frequency of the oscillations within the correlation data.

## 5.10 Future Work

In continuation of our work, we propose the application of our analysis methods to new experiments. First, we propose to replace one fish with a “dummy” decoy leader fish that would only swim laps in the annular tank. The intended effect of this would be to try eliminate non leader-follower behaviors so that the experimental data would be more similar to the model generated data. We also propose to modify the experimental setup by reducing the width of our annular tank to reduce the radial motion of the fish, enabling us to better test our hypothesis that the radial motion of our experimental fish weakens the signal of the leader follower dynamic in the angular data. Finally, we also propose to apply our metric pairwise to larger groups to investigate whether leader-follower dynamics can be identified in more complicated circumstances and if so, how their characterization differs from the pairwise case.

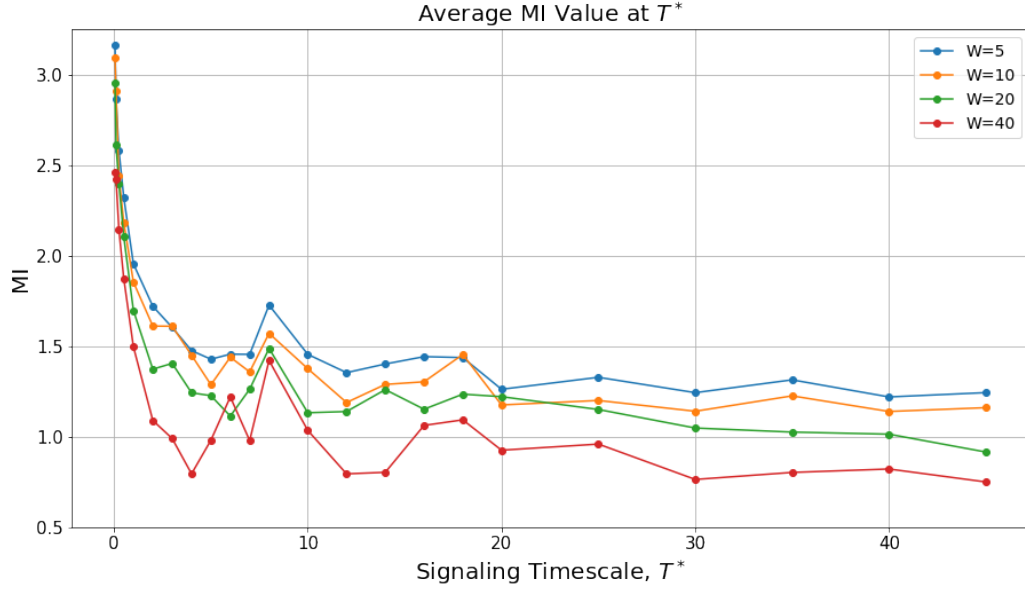


Figure 5.16: In this plot we show how the mutual information computed at the true time lag  $T^*$  varies as  $T^*$  increases. We see that the mutual information is highest for small  $T^*$ , and appears to decrease as  $T^*$  increases. The values reported are the average of  $MI(\theta_0(t); \theta_1(t + T^*))$  and  $MI(\theta_0(t); \theta_1(t - T^*))$ . Although there appears to be noise, the overall trend is decreasing.

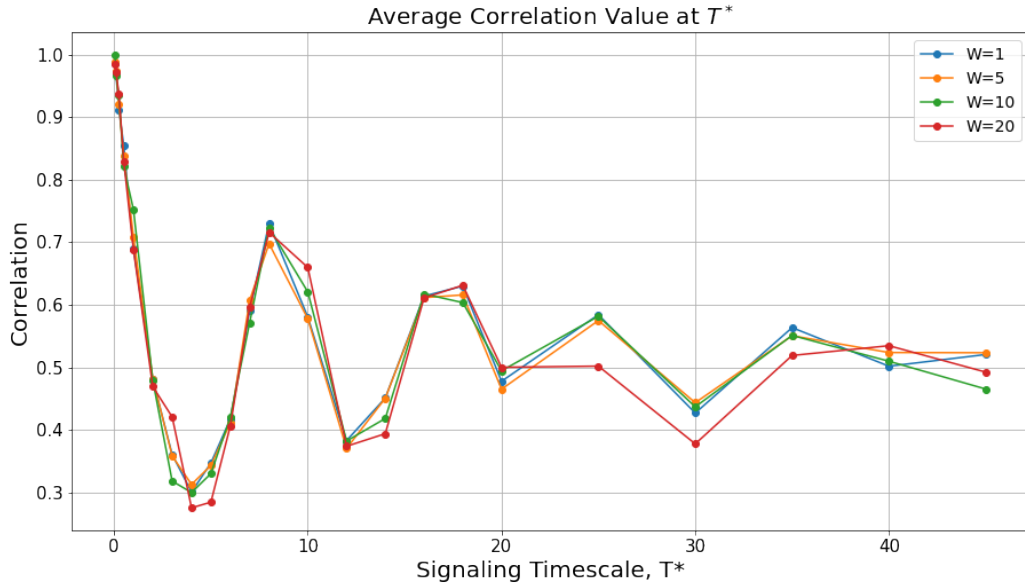


Figure 5.17: Similarly, we compute how the correlation at the true time lag  $T^*$  changes as  $T^*$  increases. The values reported are the average of  $C(\theta_0(t); \theta_1(t + T^*))$  and  $C(\theta_0(t); \theta_1(t - T^*))$ . Interestingly, despite the fact that we are considering only the correlations at the true timescale of signal transfer, the maximum value of the correlations continue to oscillate in a fashion that does not appear to be merely noise, as in the case of the mutual informations.

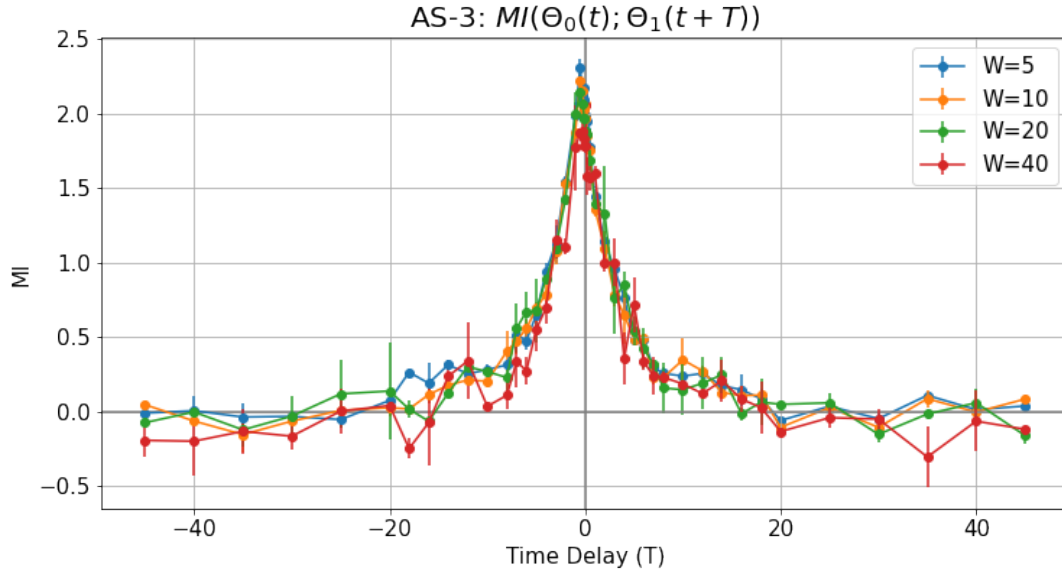


Figure 5.18: Mutual information decay for various window sizes  $W \in [5, 10, 20, 40]$  for AS-3.

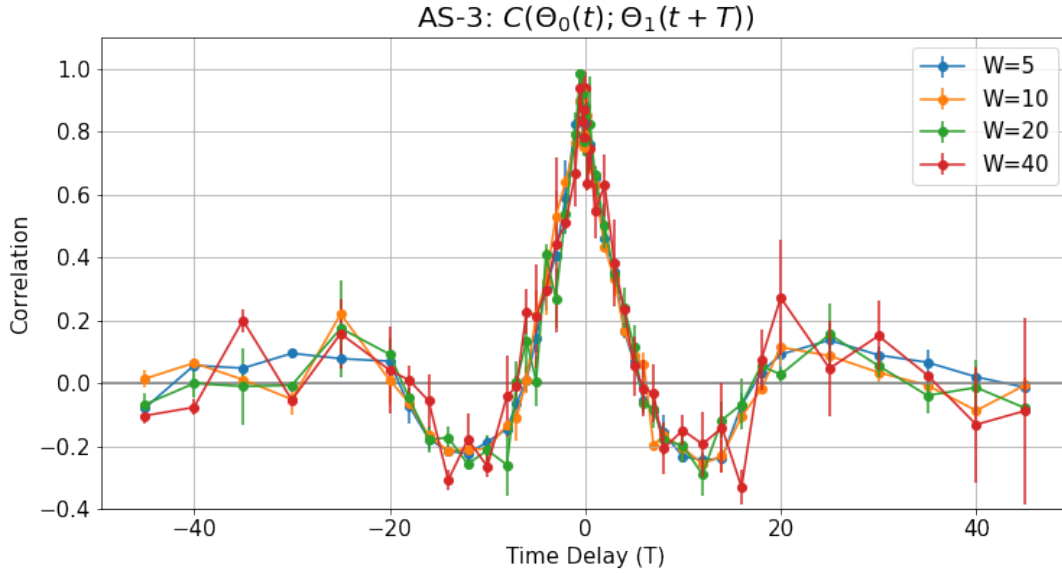


Figure 5.19: Temporal correlation plots for various window sizes  $W \in [5, 10, 20, 40]$  for AS-3.

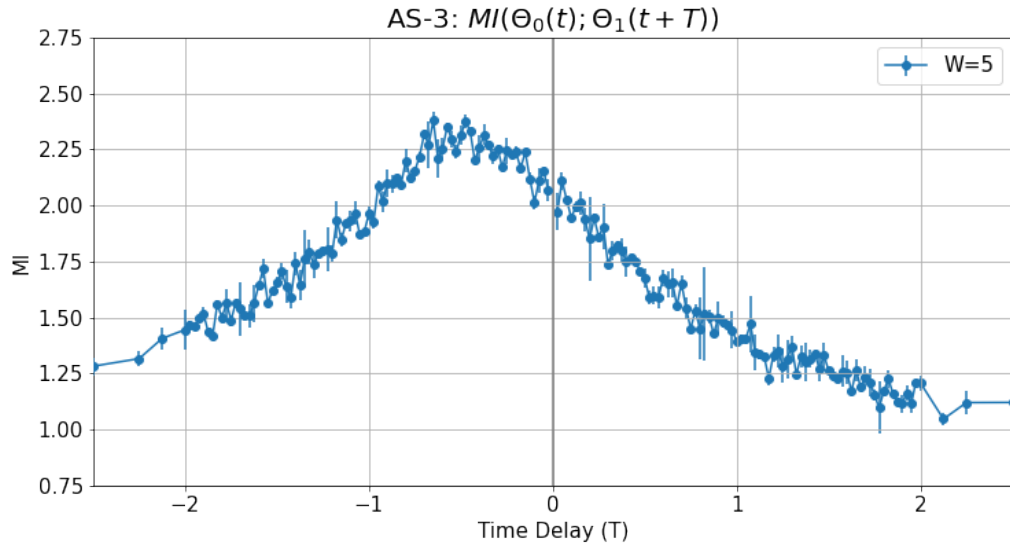


Figure 5.20: Mutual information decay for  $W = 5$  with more granularity in the reported mutual information of AS-3 for smaller values of  $T$ .

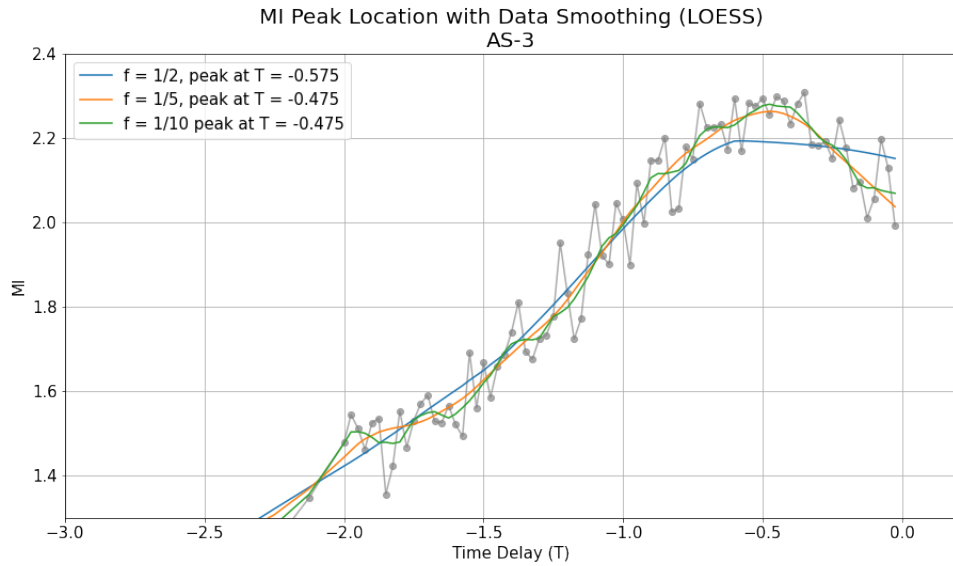


Figure 5.21: Varying the data fraction  $f$  changes the smoothness of the LOESS fit of the mutual information decay  $MI(\theta_0(t); \theta_1(t - T))$  corresponding to the negative  $T$  axis in Fig. 5.20. Larger data fractions  $f$  return smoother fits, whereas smaller fractions follow the noisiness of the data more closely.

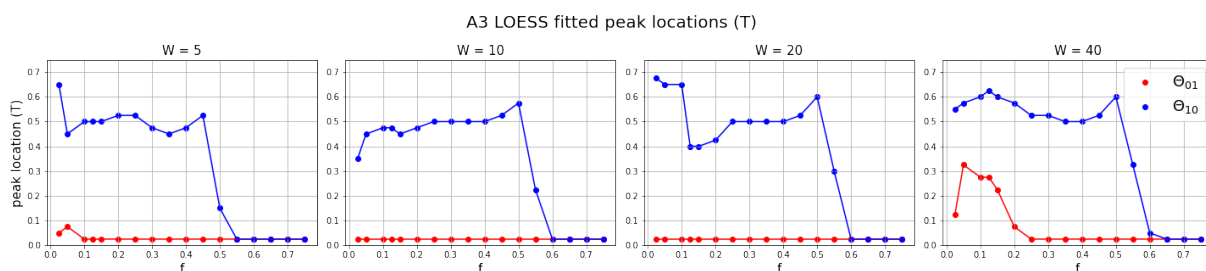
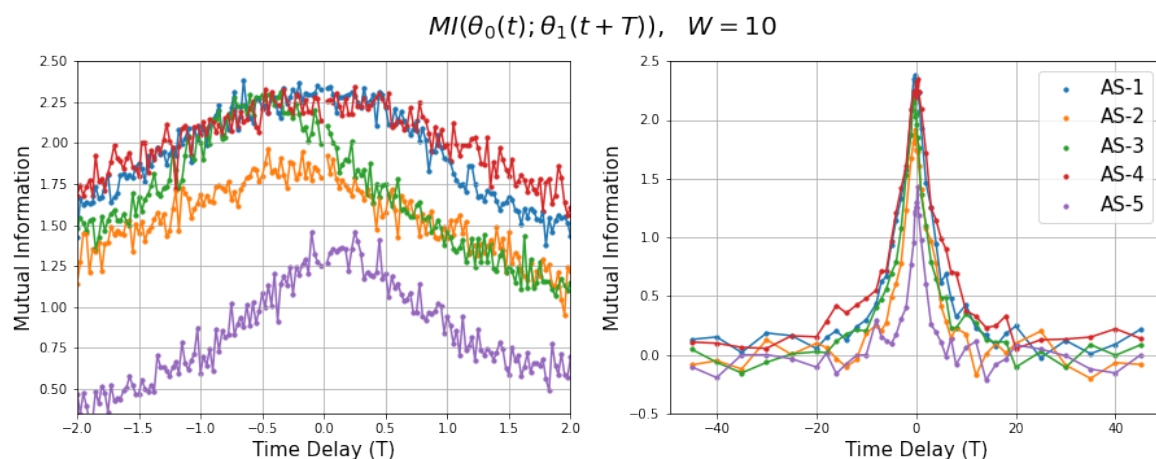
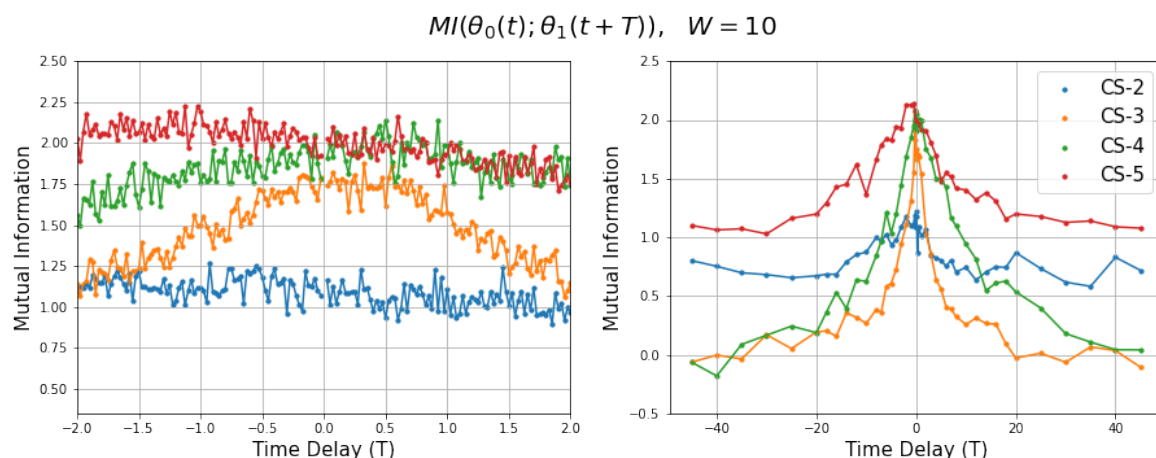


Figure 5.22: Varying the data fraction  $f$  changes the location of the maximum mutual information value, which we believe reflects the true signaling timescale ( $T^*$ ) of the experiment. Similarly, choice of separation window  $W$  changes the approximate peak location.

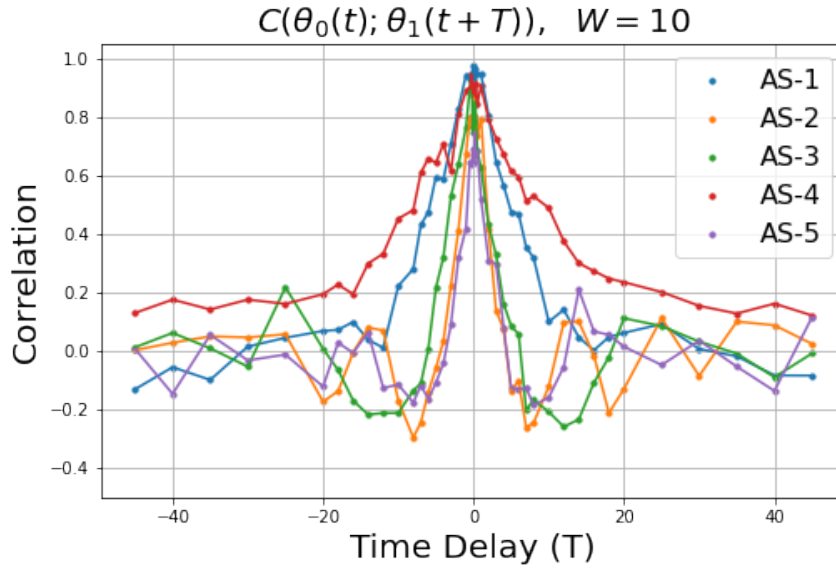


(a) Agitated MI Data

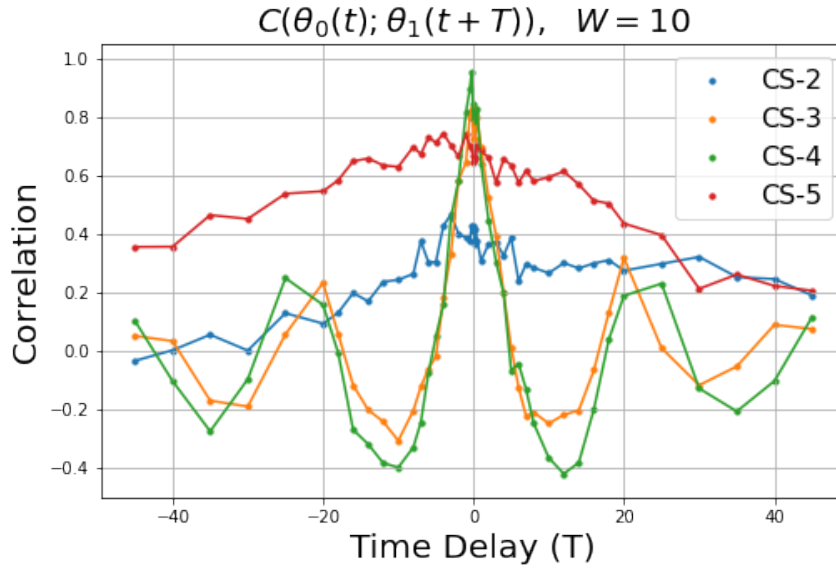


(b) Control MI Data

Figure 5.23: For a window size of  $W = 10$ , mutual information decay of the  $\theta$  variables of each experiment is shown. Left plots show the mutual information zoomed in on the  $T$  axis with higher granularity.



(a) Agitated Correlation Data



(b) Control Correlation Data

Figure 5.24: For a window size of  $W = 10$ , correlation decay of the  $\theta$  variables of each experiment is shown.

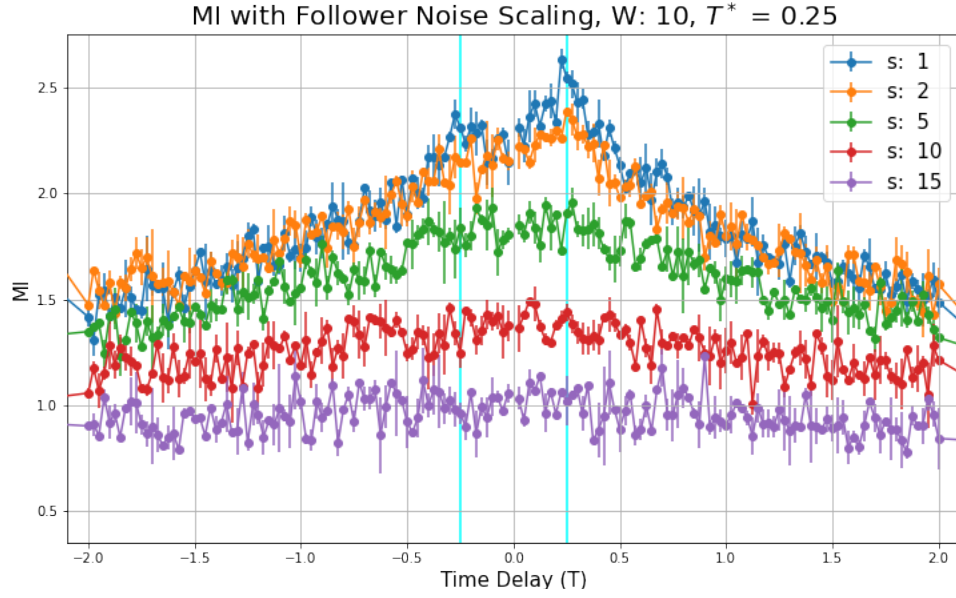


Figure 5.25: Scaling the noise by a factor of 2 or more effectively makes the peak at  $|T| = T^* = 0.25$  disappear.

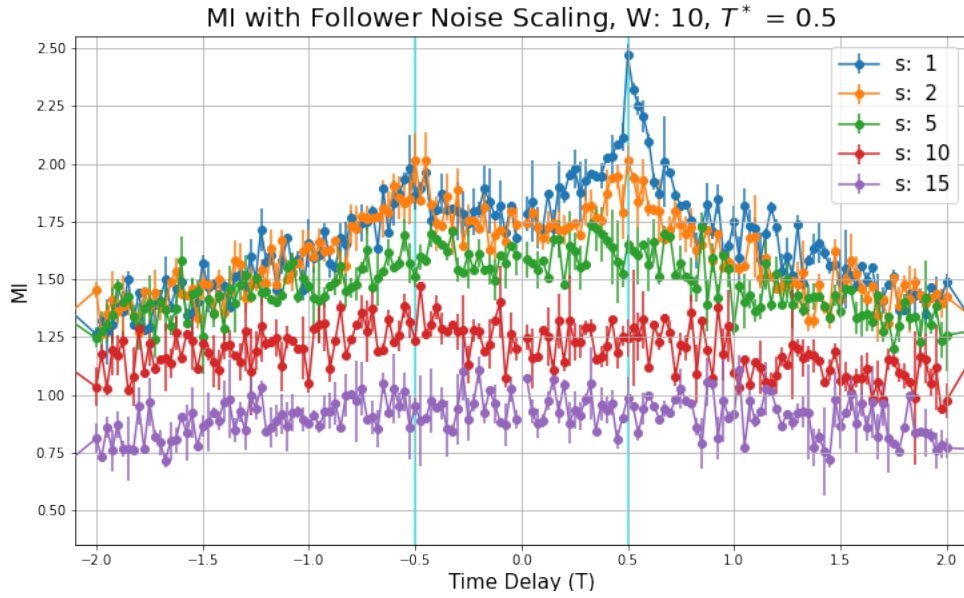


Figure 5.26: Scaling the noise by a factor of 5 or more effectively makes the peak at  $|T| = T^* = 0.5$  disappear.

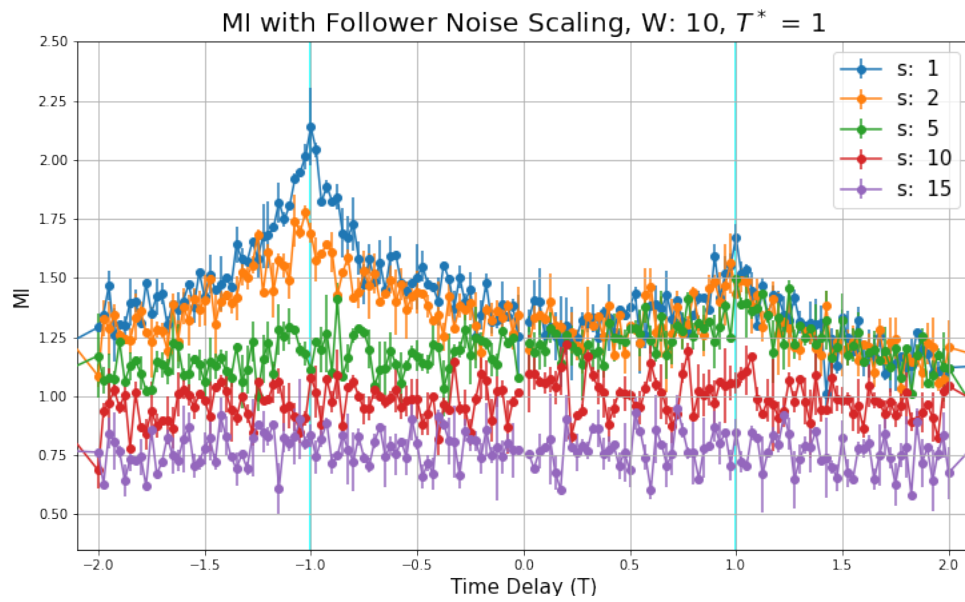


Figure 5.27: Scaling the noise by a factor of 10 or more effectively makes the peak at  $|T| = T^* = 1$  disappear.

## CHAPTER 6

### Conclusions

In this dissertation, we have explored some of the challenges that can arise when trying to understand the self and pairwise interactions in active systems using only passively gathered path data. We found that under the right conditions, the suite of current methods for analyzing interactions using path data could be improved by including the time delayed mutual information. In particular, the condition that we proposed was most important for this approach is that distribution from which the time series data is derived should be stationary. We argued that spatial confinement was more likely to produce a system that reached a stationary state and provided evidence that this was satisfied in the two systems we studied. Within this framework, we successfully estimated the time delayed mutual information from the time series data using the approach in [32] and showed that we could measure key features about our systems.

The first system that we studied was a microscale theoretical model of self avoidant swimming droplets in which the “interaction rules” were mathematically defined within the model equations. In Chapter 3, we showed the importance of using the correct model to represent the system dynamics; we found that our self-avoidant swimming droplets sometimes engaged in transient, but unpredictable self-trapping, which could not be explained by a simplified model (active Brownian motion). These unique behaviors within our model were revealed by discrepancies within regular path data analysis tools, including the mean square displacement and the velocity correlation function. We identified that the path feature associated to the emergent transient self-trapping was periods of high curvature as the particle path became spiral-like during self-trapping.

In Chapter 4, we rigorously analyzed the statistics of this path feature (curvature). First, we used the multi-scale straightness index from [55] as an order parameter to track the curvature throughout the experiment. This order parameter was itself a random variable which we called  $S(t)$  which has finite support on the interval  $[0, 1]$ . Unlike the positional data  $\mathbf{X}(t)$  which explored an infinite domain in our model, the finite support of  $S(t)$  made the distribution of  $S(t)$  at long times far more likely to be stationary (unlike  $\mathbf{X}(t)$ ). We argued that the assumption of stationary was satisfied based on the relatively constant mean and variance

of the ensemble. Using the self time-delayed mutual information on the curvature data  $S(t)$ , we found that the mutual information decayed exponentially as a function of the time delay. The rate of this exponential decay was able to detect differences in the memory expression of individual paths (shown in Fig. 4.11, across variations in the memory strength of our model at the ensemble level (Fig. 4.16), and between our model and two other models with deterministic sources of curvature (Fig. 4.17).

The second system that we studied in Chapter 5 was a macroscale active system of experimental golden shiner pairs in which the exact interaction rules were unknown; we accompanied this experiment with a reduced model in which the exact interaction rules were known. The experiments were carried out in a controlled environment in which all extraneous environmental factors were removed; this allowed us to gather data that was significantly longer than what is gathered in the wild and it removed possible confounding environmental factors. The golden shiner pairs exhibited observable behavioral symmetry as they cycled through similar behavioral modes together. There was notable positional asymmetry within these behavioral modes- one fish was in front of the other in an apparent leader-follower configuration. Unlike the swimming droplets, we were able to use the path data in our mutual information calculation since the confined experimental setup made reduced the support of the positional random variables  $\theta_0(t)$  and  $\theta_1(t)$  to the finite interval,  $[-\pi, \pi]$ . Therefore, at the long time scales which we had access to, the positional distributions of  $\theta_0(t)$  and  $\theta_1(t)$  were likely stationary. By comparing our experimental results with our theoretical model of a leader-follower pair, we estimated the signaling time scale (or reaction time),  $T^*$ , of the fish when they were in a leader-follower configuration as the peak or maximum value locations of the mutual information decay curves (reference Fig. 5.11 with Fig. 5.20 and Fig. 5.21). The extraction of the signaling timescale  $T^*$  from the experimental mutual information in agreement with our theoretical model is significant since it shows that the follower fish not passively or coincidentally following a nearby fish, but it is *actively altering its own trajectory* in response to the leader on the timescale  $T^*$ . Therefore, rather than justifying our interpretation with observations from biology, we show mathematically that there exists a leader-follower component to the social behavior of golden shiners.

In this work, we leveraged long data sets (golden shiners) and our ability to generate path data for an ensemble of agents (swimming droplets) to understand more fully the appropriate uses and conditions for using mutual information on time series data. A especially notable feature of our work is the mathematically rigorous treatment of dynamical correlations within time series data, which are overlooked in the current literature. We showed that under the right circumstances, our approach is able to capture time-dependent

nonlinear dynamics *without* conditioning on the past history of random variables (as is required for other information theoretic metrics that are designed to detect causality, including transfer entropy and causation entropy). Furthermore, we were able to accomplish this non-parametrically; the methods we developed are versatile and could be applied to any time series data that is known to sample from a stationary distribution or is likely to sample from a stationary distribution at long times due to either confinement or finite support. Given our results, there are promising future directions for the method that we have introduced, including the inclusion of other random variables into the mutual information calculation and motivation for different experiments.

## APPENDIX A: SAMPLING SCHEME AND MUTUAL INFORMATION ESTIMATION CODE

### A.1 Time Delayed Sampling Scheme

Input data must be a pandas dataframe with uniquely labeled column headers. Time must also be included as a column and it must be labeled as 't'.

```
1  def sample_independent_windows(input_data, T_start, T_end, W,
2      Timelag):
3
4      # shape and label of final data
5      num_cols = np.shape(input_data)[1]
6      headers = [*list(input_data.columns)]
7
8      # find times corresponding to all windows/intervals
9      Int_start_times = np.linspace(T_start, T_end - W,
10          int((T_end-T_start)/W )) # start, end, number of windows
11
12      # make structure to record samples
13      initial_sample = np.ones((len(Int_start_times), num_cols))*np.NaN
14      lagged_sample = np.ones((len(Int_start_times), num_cols))*np.NaN
15
16      # go through all partitions
17      for i in range(0, len(Int_start_times)):
18
19          # pick out time slice and reindex so df.sample() works
20          partition = input_data[(input_data['t'] > Int_start_times[i])
21              & (input_data['t'] < (Int_start_times[i]
22              + int(W) ))].reset_index(drop=True)
23
```

```

24     # take sample if partition is not empty
25     if(len(partition )>0):
26
27         #---take initial sample
28         sample_initial = partition.sample(n = 1, replace = False,
29             random_state = None)
30
31         # ---take timelagged sample
32         # start t
33         start = float(sample_initial['t'])
34
35         # identify end time of interval
36         target = float(sample_initial['t']) + Timelag
37
38         later_partition = input_data[(input_data['t'] >= start)
39             & (input_data['t'] <= target )].reset_index(drop=True)
40
41         #take the last item from later partition
42         new_data = later_partition.iloc[ -1 ]
43
44         initial_sample[i,:] = np.array([np.ravel(sample_initial)])
45         lagged_sample[i,:] = np.array([np.ravel(new_data)])
46
47     else:
48         i = i+1
49
50     initial_sample = pd.DataFrame(data=initial_sample,
51         columns=headers).dropna() #, index=sample.index)
52     lagged_sample = pd.DataFrame(data=lagged_sample,

```

```

53         columns=headers).dropna() #, index=sample.index)
54
55     # check to see if lagged dataframe didn't have enough rows
56     # available at the end
57     if(len(initial_sample) - len(lagged_sample) != 0):
58         # find how many extra rows
59         mismatch = len(initial_sample)- len(lagged_sample)
60
61         # drop extra rows
62         initial_sample.drop(initial_sample.tail(mismatch).index,
63                             inplace=True)
64
65     return(initial_sample, lagged_sample)
66

```

---

## A.2 KNN Estimator for Mutual Information

As noted in [32], the distance function used to compute the max will vary depending on the data type. In the code below, the data inputs are arrays of samples from the variable  $X$  and the variable  $Y$ . However, if  $X$  and  $Y$  were angles, a modified distance function would need to be used to ensure that the minimum angle between  $X$  and  $Y$  is used. (This is discussed in Chapter 5.)

```

1  def compute_MI(X, Y):
2
3      # ---- find nearest neighbor ----
4      L = len(X)
5      ni_X = []
6      ni_Y = []
7      for i in range(0, L):
8
9          # compute subspace dist between ref point and every other point

```

```

10     d = []
11     dist = np.zeros((L,2))
12     for j in range(0,L):
13         temp_x = np.abs(X[i]- X[j])
14
15         temp_y = np.abs(Y[i]- Y[j])
16
17         d.append(max(temp_x,temp_y))
18
19         dist[j,:] = [temp_x, temp_y]
20
21     d = np.sort(d)
22
23     # record distance to nearest neighbor
24     ei = d[1]
25
26     dist_x = dist[:,0]
27     dist_y = dist[:,1]
28
29     # count neighbors in subspaces
30     num_X = len(dist_x[(0 < dist_x) & (dist_x < ei)])
31     num_Y = len(dist_y[(0 < dist_y) & (dist_y < ei)])
32
33     ni_X = np.append(ni_X,num_X)
34     ni_Y = np.append(ni_Y,num_Y)
35
36     nx = ni_X + 1
37     ny = ni_Y + 1
38

```

```
39     I = digamma(K-1) - np.mean(digamma(nx) + digamma(ny)) + digamma(L)
40
41     return(I)
```

---

## APPENDIX B: ADDITIONAL CALCULATIONS FOR SWIMMING DROPLET MODEL

### B.1 Solving the Diffusion Equation to Combine and Nondimensionalize the Coupled System

Consider the dimensional, 2D system given in Eq. (3.1). Taking the Fourier Transform of Eq. (3.1a) we arrive at the ODE

$$\hat{c}_t + D|\mathbf{k}|^2 \hat{c} = \frac{\alpha D R^2}{2\pi} \exp \left[ -\frac{R^2}{2} |\mathbf{k}|^2 + i(\mathbf{k} \cdot \mathbf{X}(t)) \right]. \quad (\text{B.1})$$

We compute the integrating factor of Eq. (B.1) which is  $I = e^{\int D|\mathbf{k}|^2 dt} = e^{Dt|\mathbf{k}|^2}$ . From this Eq. (B.1) can be rewritten as

$$\frac{d}{dt} \left( e^{Dt|\mathbf{k}|^2} \hat{c} \right) = \frac{\alpha D R^2}{2\pi} e^{Dt|\mathbf{k}|^2} e^{\left[ -\frac{R^2}{2} |\mathbf{k}|^2 + i(\mathbf{k} \cdot \mathbf{X}(t)) \right]}. \quad (\text{B.2})$$

Integrating both sides of Eq. (B.2) gives the solution to Eq. (B.1)

$$\hat{c} = \frac{\alpha D R^2}{2\pi} \int_0^t e^{-D(t-s)|\mathbf{k}|^2} e^{\left[ -\frac{R^2}{2} |\mathbf{k}|^2 + i(\mathbf{k} \cdot \mathbf{X}(s)) \right]} ds. \quad (\text{B.3})$$

Taking the inverse Fourier Transform of Eq. (B.3) yields the solution,  $c(\mathbf{x}, t)$ , to Eq. (3.1a), which is

$$c = \frac{\alpha D R^2}{2\pi} \int_0^t (R^2 + 2D(t-s))^{-1} \cdot e^{-\frac{|\mathbf{x} - \mathbf{X}(s)|^2}{2(R^2 + 2D(t-s))}} ds. \quad (\text{B.4})$$

We can incorporate the solution to Eq. (3.1a), which is Eq. (B.4), into Eq. (3.1b) by taking the gradient,  $\nabla c$ , which is

$$\nabla c = -\frac{\alpha D R^2}{2\pi} \int_0^t (R^2 + 2D(t-s))^{-2} (\mathbf{x} - \mathbf{X}(s)) \cdot e^{-\frac{|\mathbf{x} - \mathbf{X}(s)|^2}{2(R^2 + 2D(t-s))}} ds. \quad (\text{B.5})$$

The SDE path evolution Eq. (3.1b) then becomes

$$d\mathbf{X} = \frac{\alpha D \beta R}{(2\pi)^2} \left[ \int_0^t (R^2 + 2D(t-s))^{-2} \left( \int_{\mathbb{R}^2} (\mathbf{x} - \mathbf{X}(s)) \cdot e^{-\frac{|\mathbf{x} - \mathbf{X}(s)|^2}{2(R^2 + 2D(t-s))} - \frac{|\mathbf{x} - \mathbf{X}(t)|^2}{2R^2}} d\mathbf{x} \right) ds \right] dt + \sqrt{\sigma} d\mathbf{W}. \quad (\text{B.6})$$

Evaluation of the spatial integral over  $\mathbb{R}^2$  reduces Eq. (B.6) to

$$d\mathbf{X} = \frac{\alpha D \beta R^3}{2^3 \pi} \left[ \int_0^t \left( (\mathbf{X}(t) - \mathbf{X}(s)) e^{-\frac{|\mathbf{X}(t) - \mathbf{X}(s)|^2}{4(R^2 + D(t-s))}} (R^2 + D(t-s))^{-2} \right) ds \right] dt + \sqrt{\sigma} d\mathbf{W}. \quad (\text{B.7})$$

By nondimensionalizing under the scalings  $\mathbf{Y} = \frac{\mathbf{X}}{R}$ ,  $\tau = \frac{t}{T}$ , and  $\mathbf{B} = \frac{\mathbf{W}}{\sqrt{T}}$ , Eq. (B.7) becomes

$$Rd\mathbf{Y} = \frac{\alpha D \beta R^3}{2^3 \pi} \left[ \int_0^\tau \left( e^{-\frac{|R\mathbf{Y}(\tau) - R\mathbf{Y}(\zeta)|^2}{4(R^2 + DT(\tau - \zeta))}} (R\mathbf{Y}(\tau) - R\mathbf{Y}(\zeta))(R^2 + DT(\tau - \zeta))^{-2} \right) T d\zeta \right] T d\tau + \sqrt{\sigma T} d\mathbf{B}. \quad (\text{B.8})$$

The SDE path evolution given by Eq. (B.8) then simplifies to

$$d\mathbf{Y} = \frac{\alpha D \beta R^3 T^2}{2^3 \pi} \left[ \int_0^\tau \left( e^{-\frac{|R\mathbf{Y}(\tau) - R\mathbf{Y}(\zeta)|^2}{4(R^2 + DT(\tau - \zeta))}} (\mathbf{Y}(\tau) - \mathbf{Y}(\zeta))(R^2 + DT(\tau - \zeta))^{-2} \right) d\zeta \right] d\tau + \frac{\sqrt{\sigma T}}{R} d\mathbf{B} \quad (\text{B.9})$$

Incorporating the nondimensional parameters  $D \rightarrow \mu = \frac{DT}{R^2}$ ,  $\alpha \rightarrow \phi = \frac{\alpha R^2}{2\pi}$ ,  $\beta \rightarrow \nu = \frac{\beta T}{2\pi R}$ , and  $\sigma \rightarrow \epsilon = \frac{\sigma T}{R^2}$  and exchanging  $s$  for  $\zeta$  and  $t$  for  $\tau$  for notational convenience we have the nondimensional SDE path evolution equation

$$d\mathbf{Y} = \frac{\pi}{2} \mu \nu \phi \left[ \int_0^t \left( e^{-\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(1 + \mu(t - s))}} (\mathbf{Y}(t) - \mathbf{Y}(s))(1 + \mu(t - s))^{-2} \right) ds \right] dt + \sqrt{\epsilon} d\mathbf{B} \quad (\text{B.10})$$

in agreement with Eq. (3.3).

## B.2 Computation of the Velocity Integral Formulation Using a Dirac Delta Function

To assess the case in which the particle is considered a point source, we substitute the mollified delta function,  $\delta_R(\mathbf{x} - \mathbf{X}(t)) = \frac{1}{2\pi R^2} e^{-\frac{|\mathbf{x} - \mathbf{X}(t)|^2}{2R^2}}$ , in Eq. (3.2) for a Dirac delta function,

$$\frac{\partial c}{\partial t} = D\Delta c + \alpha D \delta^2(\mathbf{x} - \mathbf{X}(t)) \quad (\text{B.11a})$$

$$d\mathbf{X}(t) = -\beta R \left( \int_{\Omega} \delta^2(\mathbf{x} - \mathbf{X}(t)) \nabla c d\mathbf{x} \right) dt + \sqrt{\sigma} d\mathbf{W}. \quad (\text{B.11b})$$

Here,  $\delta^2(\mathbf{x} - \mathbf{X}(t))$  is a 2-dimensional Dirac delta function centered at  $\mathbf{X}(t)$ . The  $R^2$  in the source term of the original PDE given by Eq. (3.1a) is no longer necessary. Accordingly, the units of  $\alpha$  are  $[\alpha] = c$  and the units of  $\beta$  remain  $[\beta] = \frac{L}{cT}$ . Nondimensionalizing Eq. (B.11) with the scalings  $\mathbf{y} = \frac{\mathbf{x}}{R}$ ,  $\mathbf{Y} = \frac{\mathbf{X}}{R}$ ,  $\tau = \frac{t}{T}$ , and  $\mathbf{B} = \frac{\mathbf{W}}{\sqrt{T}}$  and where  $\mu = \frac{DT}{R^2}$ ,  $\phi = \frac{\alpha}{2\pi}$ ,  $\nu = \frac{\beta T}{R^2}$  and  $\epsilon = \frac{\sigma T}{R^2}$  we arrive at the new system

$$\frac{\partial c}{\partial t} = \mu \Delta c + 2\pi \mu \phi \delta^2(\mathbf{y} - \mathbf{Y}(t)) \quad (\text{B.12a})$$

$$d\mathbf{Y}(t) = -2\pi\nu \left( \int_{\Omega} \delta^2(\mathbf{y} - \mathbf{Y}(t)) \nabla c d\mathbf{y} \right) dt + \sqrt{\epsilon} d\mathbf{B} \quad (\text{B.12b})$$

where  $c$ ,  $t$  and  $\Omega$  are re-used for their non-dimensional versions for convenience.

As in the case with the sized particle, we take the Fourier Transform of the Eq. (B.12a) to arrive at the ODE

$$\hat{c}_t + \mu|\mathbf{k}|^2 \hat{c} = \mu\phi e^{i\mathbf{k} \cdot \mathbf{Y}(t)}. \quad (\text{B.13})$$

We compute the integrating factor of Eq. (B.13) which is  $I = e^{\int \mu|\mathbf{k}|^2 dt} = e^{\mu t|\mathbf{k}|^2}$ . From this, Eq. (B.13) can be rewritten as

$$\frac{d}{dt} \left( \hat{c} \cdot e^{\mu t|\mathbf{k}|^2} \right) = \mu\phi e^{i\mathbf{k} \cdot \mathbf{Y}(t)} \cdot e^{\mu t|\mathbf{k}|^2}. \quad (\text{B.14})$$

Integrating both sides of Eq. (B.14) gives

$$\hat{c} = \mu\phi \int_0^t e^{-\mu(t-s)|\mathbf{k}|^2 + i\mathbf{k} \cdot \mathbf{Y}(s)} ds. \quad (\text{B.15})$$

We take the inverse Fourier Transform of Eq. (B.15) to find the solution to Eq. (B.12a), which is

$$c = \mu\phi \int_0^t (2\mu(t-s))^{-1} e^{-\frac{|\mathbf{y} - \mathbf{Y}(s)|^2}{4(\mu(t-s))}} ds. \quad (\text{B.16})$$

We incorporate the solution to Eq. (B.12a) into Eq. (B.12b) by computing the gradient  $\nabla c$  of Eq. (B.16), which is

$$\nabla c = -\mu\phi \int_0^t \frac{(\mathbf{y} - \mathbf{Y}(s))}{(2\mu(t-s))^2} \exp \left[ -\frac{|\mathbf{y} - \mathbf{Y}(s)|^2}{4(\mu(t-s))} \right] ds. \quad (\text{B.17})$$

Eq. B.12b then becomes

$$d\mathbf{Y}(t) = \nu\mu\phi \frac{\pi}{2} \int_0^t (\mu(t-s))^{-2} \int_{\Omega} \delta^2(\mathbf{y} - \mathbf{Y}(t)) (\mathbf{y} - \mathbf{Y}(s)) \exp \left[ -\frac{|\mathbf{y} - \mathbf{Y}(s)|^2}{4(\mu(t-s))} \right] d\mathbf{y} ds dt + \sqrt{\epsilon} d\mathbf{B}. \quad (\text{B.18})$$

Evaluation of the spatial integral over  $\mathbb{R}^2$  reduces Eq. (B.18) to

$$d\mathbf{Y}(t) = \nu\mu\phi \frac{\pi}{2} \int_0^t (\mu(t-s))^{-2} (\mathbf{Y}(t) - \mathbf{Y}(s)) \exp \left[ -\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(\mu(t-s))} \right] ds dt + \sqrt{\epsilon} d\mathbf{B}. \quad (\text{B.19})$$

Now, suppose that:  $\mathbf{Y}(t) = \langle Vt, 0 \rangle$ . This simplifies Eq. (B.19) to

$$\frac{d\mathbf{Y}(t)}{dt} = V = \nu\mu\phi\frac{\pi}{2}\int_{-\infty}^t (\mu(t-s))^{-2}(Vt - Vs) \exp\left[-\frac{|Vt - Vs|^2}{4(\mu(t-s))}\right] ds. \quad (\text{B.20})$$

By making the change of variables given by  $z = \mu(t - s)$  and  $ds = -\frac{1}{\mu}dz$ , we see that Eq. (B.20) is considerably reduced to

$$V = \nu\mu\phi\frac{\pi}{2}\int_0^\infty z^{-1}\frac{V}{\mu}\exp\left[-\left(\frac{V}{\mu}\right)^2\frac{z}{4}\right] dz. \quad (\text{B.21})$$

After some further simplification we arrive at the following expression,

$$1 = \frac{\pi}{2}\frac{\nu}{\mu}\phi\int_0^\infty \frac{1}{z}\exp\left[-\left(\frac{V}{\mu}\right)^2\frac{z}{4}\right] \frac{1}{\mu}dz. \quad (\text{B.22})$$

This integral on the right hand side is not pointwise convergent for finite  $V$  and thus indicates that when we consider the particle to be a point source with the self-avoidant memory that we have defined, the particle does not swim.

### B.3 Computation of the Hover Height Integral Formulation

This calculation is included for completeness, since this work is reproduced from [15]. In [15], the calculation was not completed by the author of this dissertation. To show that the presented model also reproduces the experimentally-observed hovering of the droplets above the bottom place, we set the second component of the position  $\mathbf{Y}$  in the direction perpendicular to the bottom plate, and add a constant non-dimensional gravitational force  $f_g$ . We then seek a steady state solution of the form  $\mathbf{Y} = (0, h)$  for non-dimensional hover height  $h$  of the droplet's center with reflecting boundary condition for the concentration field at  $\mathbf{y} = (x_1, 0)$  with  $x_1 \in \mathbb{R}$ . Using the model formulation in Eq. (3.3) we can account for this boundary condition using the standard trick of placing an image particle at  $\mathbf{Y}^* = (x_1, -x_2)$ . The resulting equation for the position  $\mathbf{Y}$  including the image particle  $\mathbf{Y}^*$  and the gravitational force is

$$\begin{aligned} d\mathbf{Y} = & \frac{\pi}{2}\mu\nu\phi\int_0^t \exp\left[-\frac{|\mathbf{Y}(t) - \mathbf{Y}(s)|^2}{4(1 + \mu(t-s))}\right] (1 + \mu(t-s))^{-2}(\mathbf{Y}(t) - \mathbf{Y}(s))dsdt \\ & + \frac{\pi}{2}\mu\nu\phi\int_0^t \exp\left[-\frac{|\mathbf{Y}(t) - \mathbf{Y}^*(s)|^2}{4(1 + \mu(t-s))}\right] (1 + \mu(t-s))^{-2}(\mathbf{Y}(t) - \mathbf{Y}^*(s))dsdt - (0, f_g) + \sqrt{\epsilon}d\mathbf{B}. \end{aligned} \quad (\text{B.23})$$

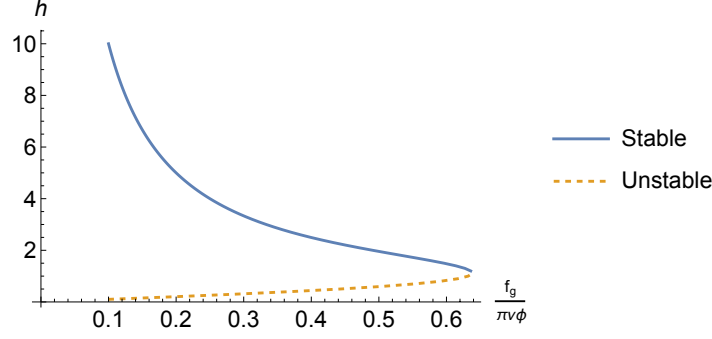


Figure B.1: Solutions to Eq. (B.25) for the hover-height  $h$  of the droplet's center above the bottom plate as a function of  $f_g/\pi\nu\phi$ . Beyond a critical value of  $f_g/\pi\nu\phi$  (corresponding to  $h \approx 1$ ) the droplets no longer hover but sit on the bottom plate.

Isolating the second component, and looking for solutions  $\mathbf{Y} = (0, h)$  and  $\mathbf{Y}^* = (0, -h)$  for all time, with no noise ( $\epsilon = 0$ ) we arrive at

$$f_g = \pi\mu\nu\phi \int_{-\infty}^t \exp \left[ -\frac{h^2}{(1 + \mu(t-s))} \right] (1 + \mu(t-s))^{-2} h ds. \quad (\text{B.24})$$

Under the change of variables  $z = \mu(t-s)$ , the above is equivalent to

$$f_g = \pi\nu\phi \int_0^\infty \exp \left[ -\frac{h^2}{1+z} \right] \frac{h}{(1+z)^2} dz \quad (\text{B.25})$$

which is independent of the memory timescale  $\mu^{-1}$  as one might intuitively expect.

Numerically-determined solutions to Eq. (B.25) as a function of  $f_g/\pi\nu\phi$  are shown in Fig. B.1. Beyond a critical value of this parameter grouping, the droplets would no longer hover and rather fall to the bottom. Note this occurs at about  $h = 1$  which is the non-dimensional radius  $R$ ; the unstable solutions are within the fictitious boundary of the droplets. A qualitative comparison to the experimental results of Fig. 3 in Ref. [44] reveals two similar trends. First, increased SDS concentration yields a higher hover height. In our model, this roughly corresponds to a stronger response to the concentration gradient, or the parameter  $\nu$ . Increasing  $\nu$  similarly increases the hover height. Second, increased radius of the particles decreased the hover height. In our model, this roughly corresponds to increasing the non-dimensional gravitational force  $f_g$  which too decreases the hover height.

#### B.4 Computing the Small Time Asymptotics of the Active Brownian MSD

Recall the MSD given for the active Brownian particle (ABP) model with translational noise and rotational diffusion given in Eq. (3.6):

$$\mathbb{E}[\mathbf{X}(t)^2] = 4V^2\tau^2 \left[ 2 \left( e^{-\frac{t}{2\tau}} - 1 \right) + \frac{t}{\tau} \right] + 2\epsilon t \quad (\text{B.26})$$

Starting from the MSD in Eq. (3.7) for the ABP model, we rewrite the exponential as an infinite series to arrive at

$$\mathbb{E}[\mathbf{X}(t)^2] = 4V^2\tau^2 \left[ 2 \left( \sum_{n=0}^{\infty} \frac{1}{n!} \left( -\frac{t}{2\tau} \right)^n - 1 \right) + \frac{t}{\tau} \right] + 2\epsilon t. \quad (\text{B.27})$$

This is asymptotic to

$$\mathbb{E}[\mathbf{X}(t)^2] \approx 4V^2\tau^2 \left[ 2 \left( \left( 1 - \frac{t}{2\tau} + \frac{t^2}{8\tau^2} \right) - 1 \right) + \frac{t}{\tau} \right] + 2\epsilon t \quad (\text{B.28})$$

as  $t \rightarrow 0$  by just retaining a few leading order terms.

In the small time scale regime where  $t^n \gg t^{n+1}$ , we see that

$$\mathbb{E}[\mathbf{X}(t)^2] \approx V^2 t^2 + 2\epsilon t \quad (\text{B.29})$$

we obtain Eq. (3.10). This expression is dominated by the diffusion-generated term  $2\epsilon t$  at the smallest time scales (where  $t \gg t^2$ ) and dominated by the directed motion term  $V^2 t^2$  when  $t^2$  becomes sufficiently larger than  $t$ .

Returning to Eq. (3.7) in the large timescale regime where  $t \gg \tau$ , we see that

$$e^{-\frac{t}{2\tau}} \rightarrow 0$$

and therefore

$$\mathbb{E}[\mathbf{X}(t)^2] \approx (4V^2\tau + 2\epsilon)t$$

as given by Eq. (3.11). This expression contains the amount of enhanced diffusion,  $4V^2\tau^2$ .

## B.5 Computing MSD and OCF from Position Time Series Generated by the Model

Absent a closed form expression for the mean square displacement of our model, we compute the empirical MSD from the position time series of length  $N + 1$  given by  $\mathbf{X}(t): \{\mathbf{X}(0), \dots, \mathbf{X}(N)\}$ . To avoid introducing any correlations into the increment averages, we use non-overlapping increments. To achieve statistical accuracy, we then average over many simulated trajectories. We denote the integer lag time as  $\Delta L$  indicating the displacement traveled by the particle between observations  $j$  and  $j + \Delta L$  and given by  $\mathbf{X}(j + \Delta L) - \mathbf{X}(j)$ . The total number of non-overlapping increments of length  $\Delta L$  in a time series of length  $N + 1$  is  $k = \lfloor \frac{N+1}{\Delta L} \rfloor$ . (In the event that the index lag length  $\Delta L$  does not evenly divide the number of increments  $N + 1$ , we remove the extra data from the beginning of the time.) Thus, the empirical formula for the mean square displacement over the lag time  $\Delta L$  of a single particle is given by

$$\Delta L^2 = \frac{1}{k-1} \sum_{i=1}^k (\mathbf{X}(N - (i-1) \cdot \Delta L) - \mathbf{X}(N - i \cdot \Delta L))^2. \quad (\text{B.30})$$

As shown in Fig. B.2, successive increases in  $\Delta L$  result in a sampling process which coarse grains the position time series.

Using the same partitioning process described above and shown in Fig. B.2 we can compute the non-overlapping displacements and find the cosine between consecutive pairs. The resulting time average of these computed cosines gives the orientation correlation function, for which the formula is given in Eq. (3.12).

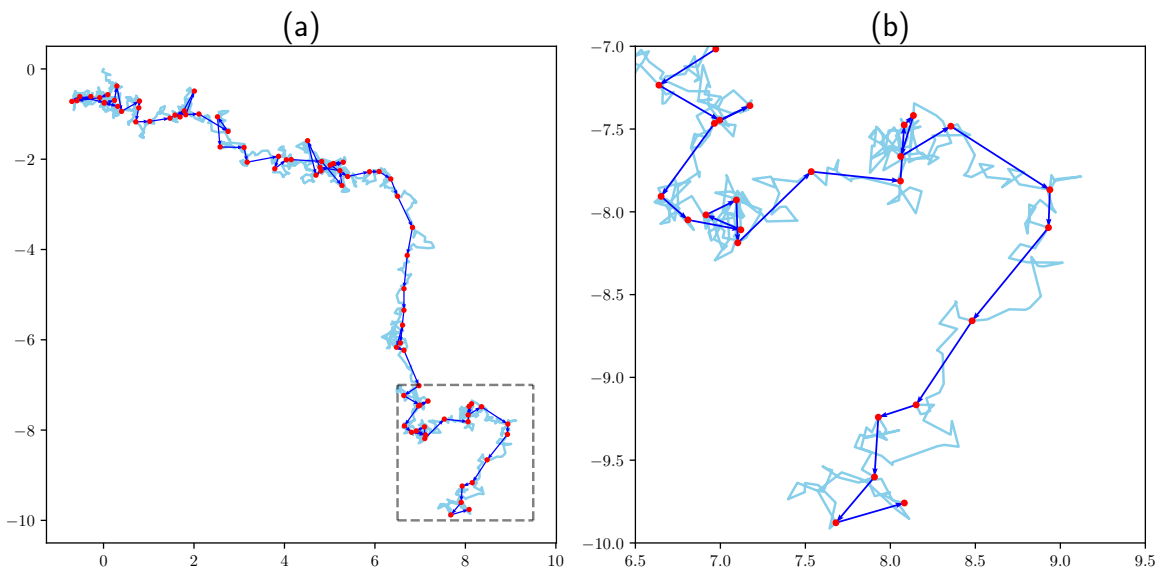


Figure B.2: Position time series of a sample trajectory with coarse grained lag times of 10 increments. Panel (a) includes entire trajectory and panel (b) is the inset identified with the dashed square.

## APPENDIX C: SUPPLEMENTAL FIGURES TO GOLDEN SHINER BEHAVIORAL ANALYSIS

### C.1 Supplemental Figures to Experimental Golden Shiner Behavioral Analysis

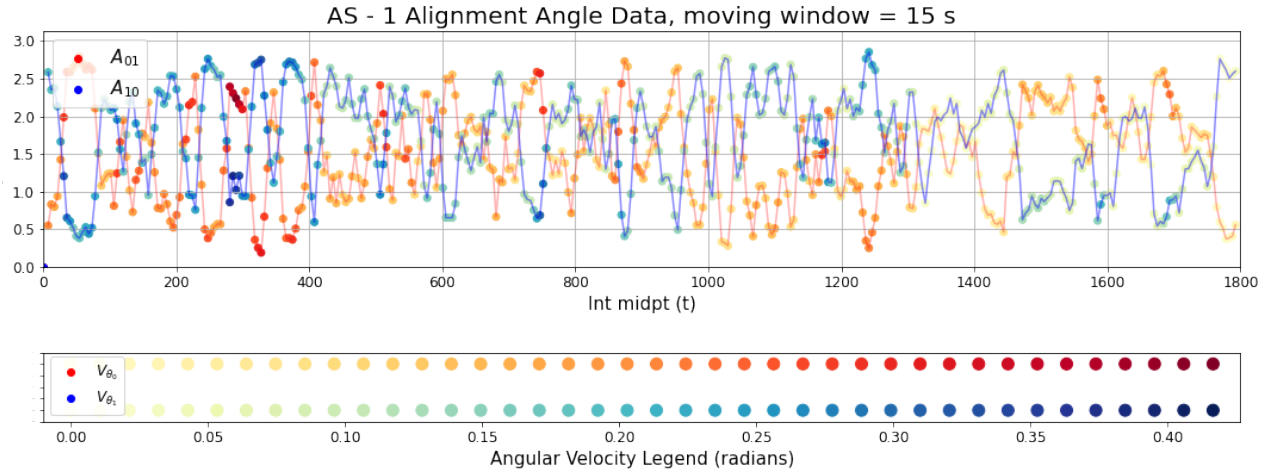


Figure C.1: Alignment data of AS-1.

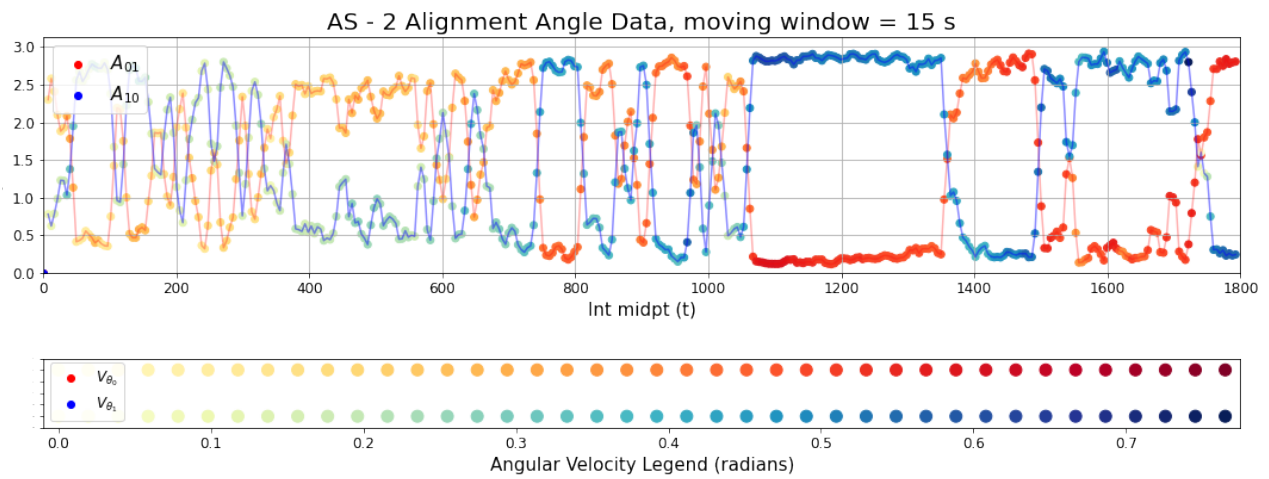


Figure C.2: Alignment data of AS-2.

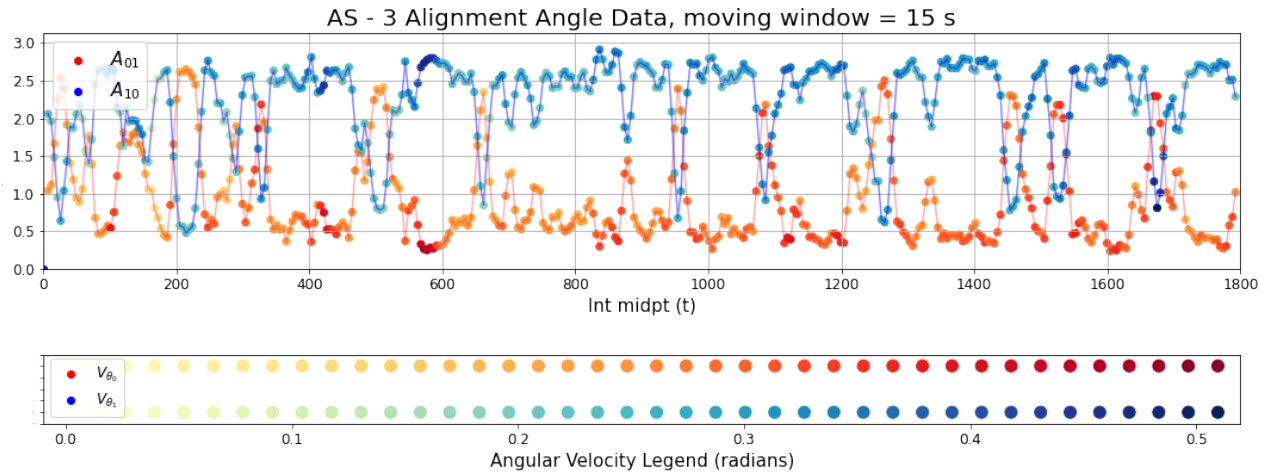


Figure C.3: Alignment data of AS-3.

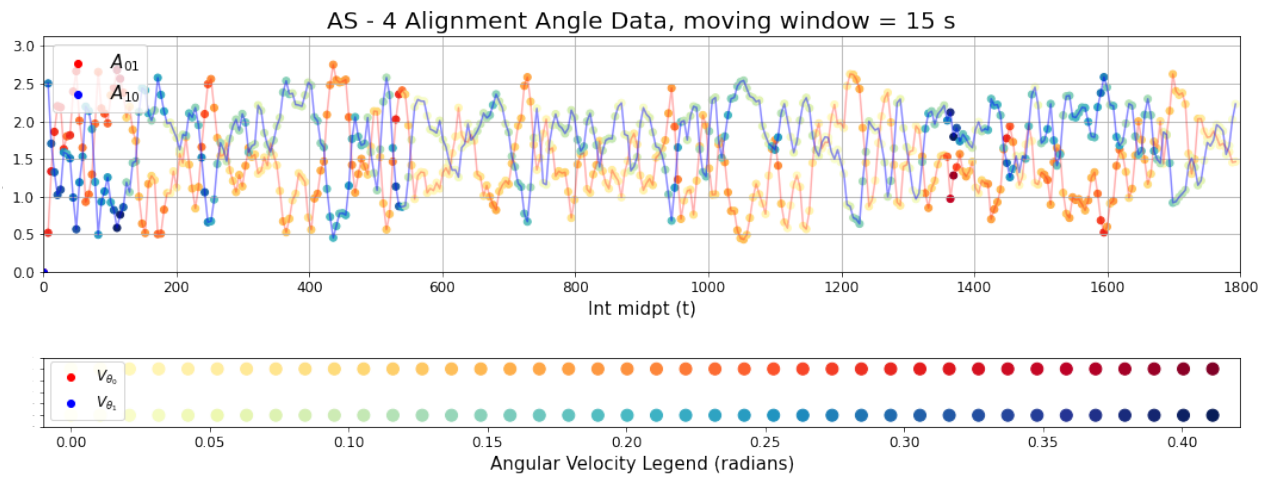


Figure C.4: Alignment data of AS-4.

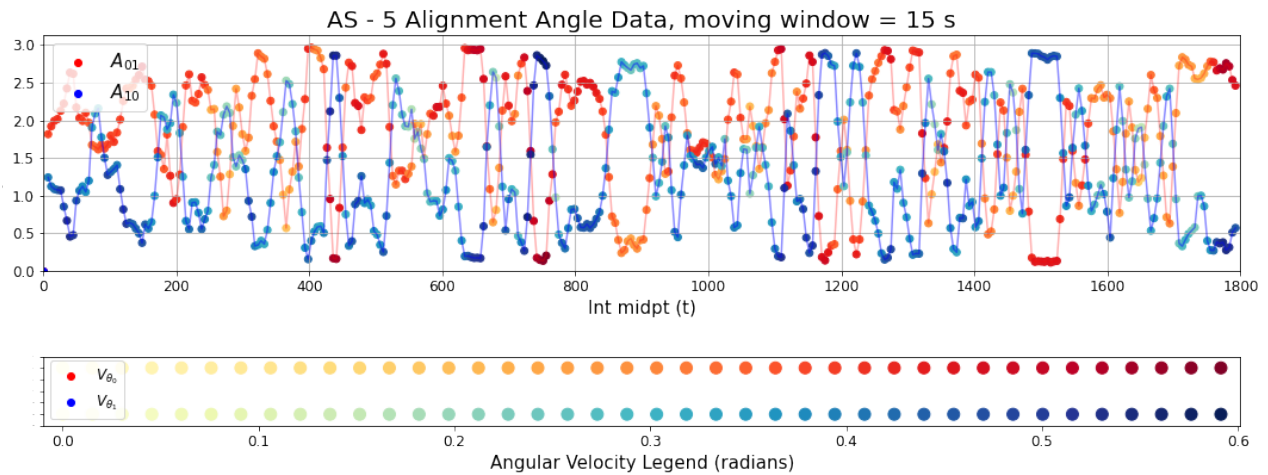


Figure C.5: Alignment data of AS-5.

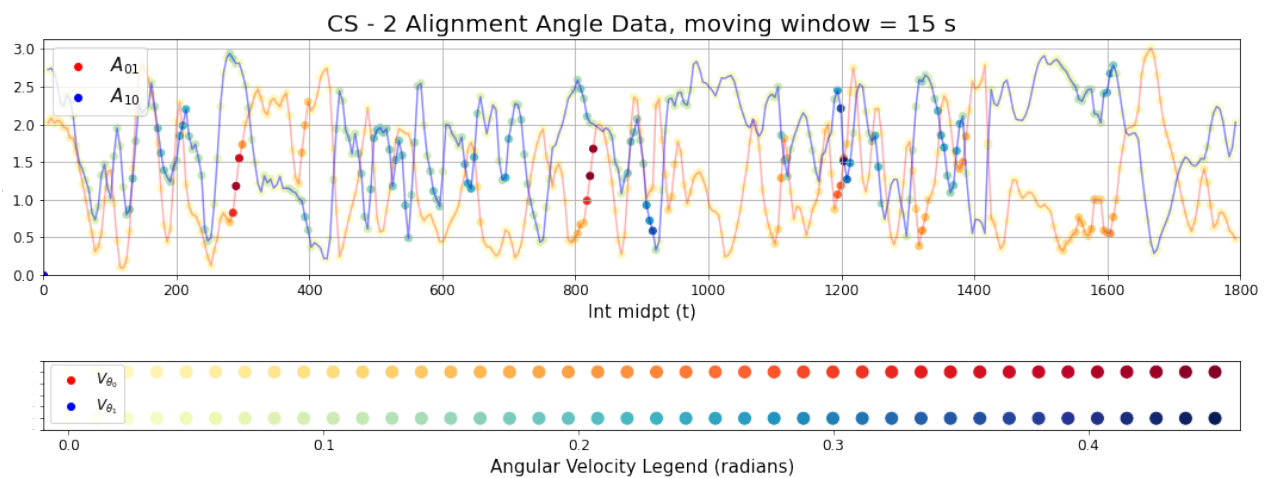


Figure C.6: Alignment data of CS-2.

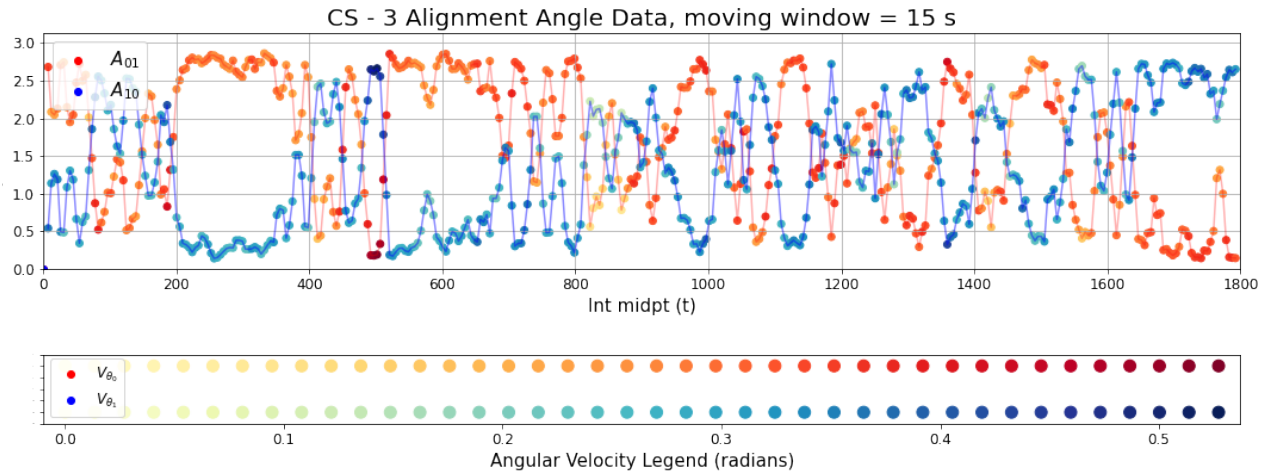


Figure C.7: Alignment data of CS-3.

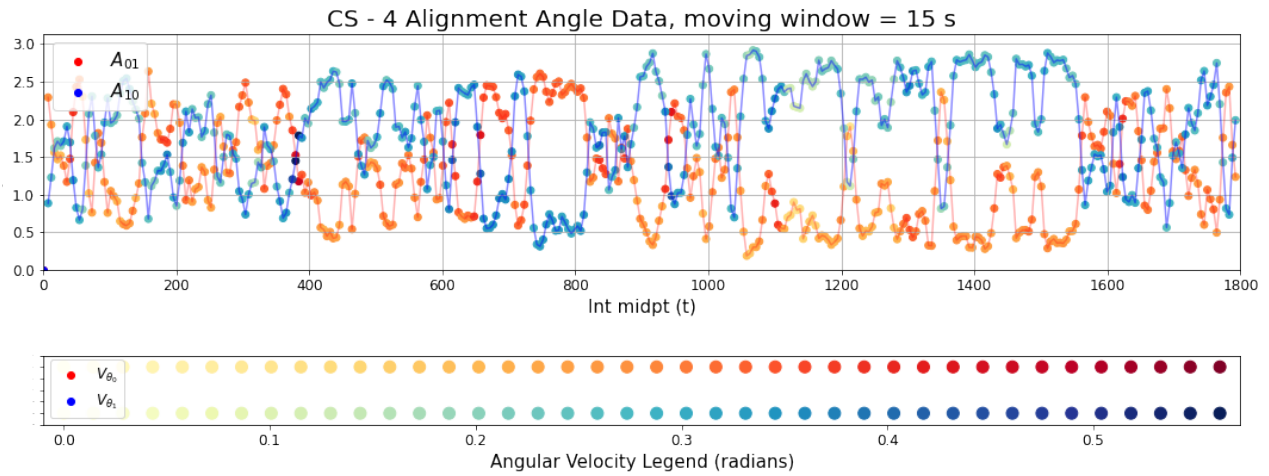


Figure C.8: Alignment data of CS-4.

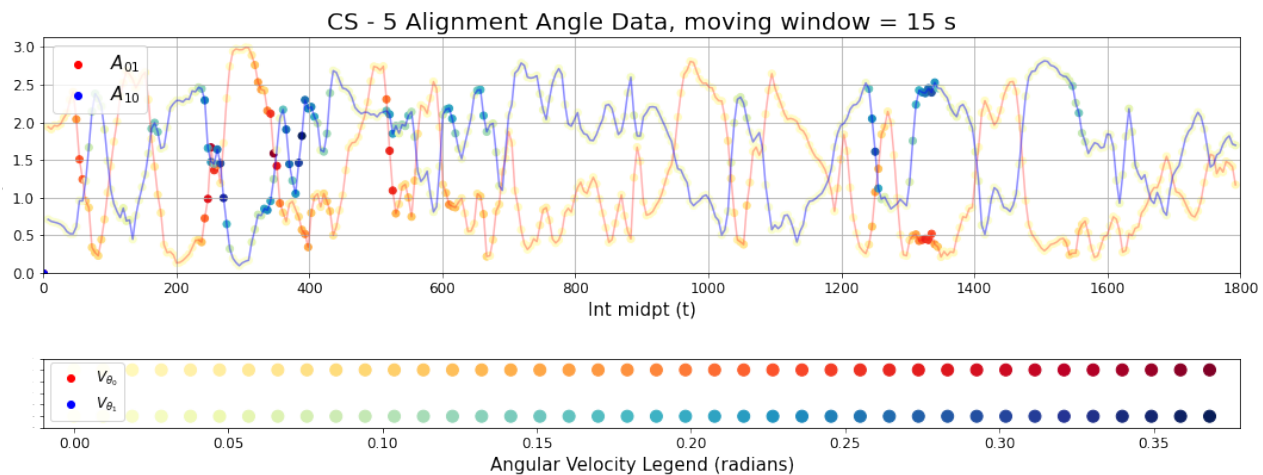


Figure C.9: Alignment data of CS-5.

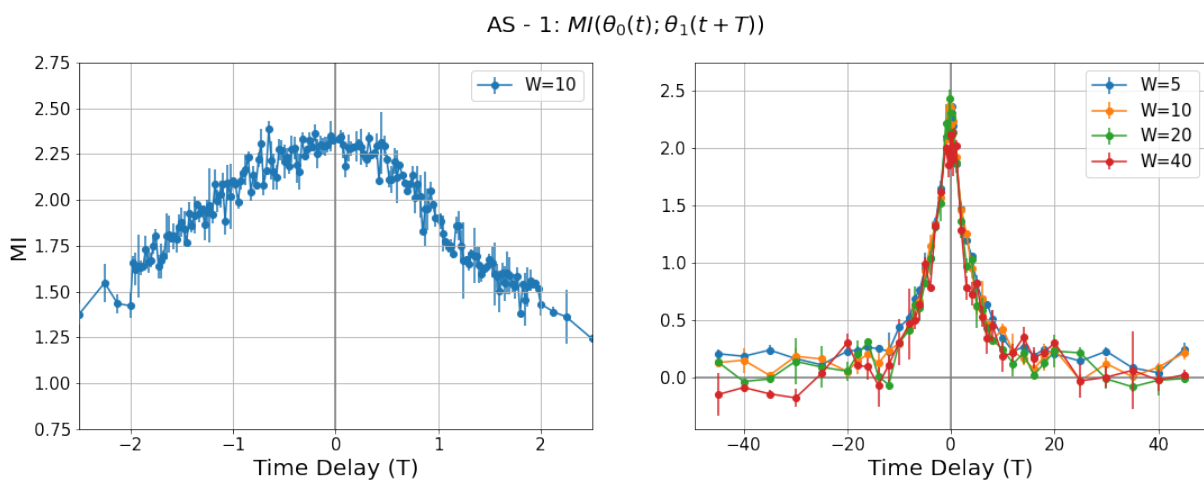


Figure C.10: Mutual information of AS-1.

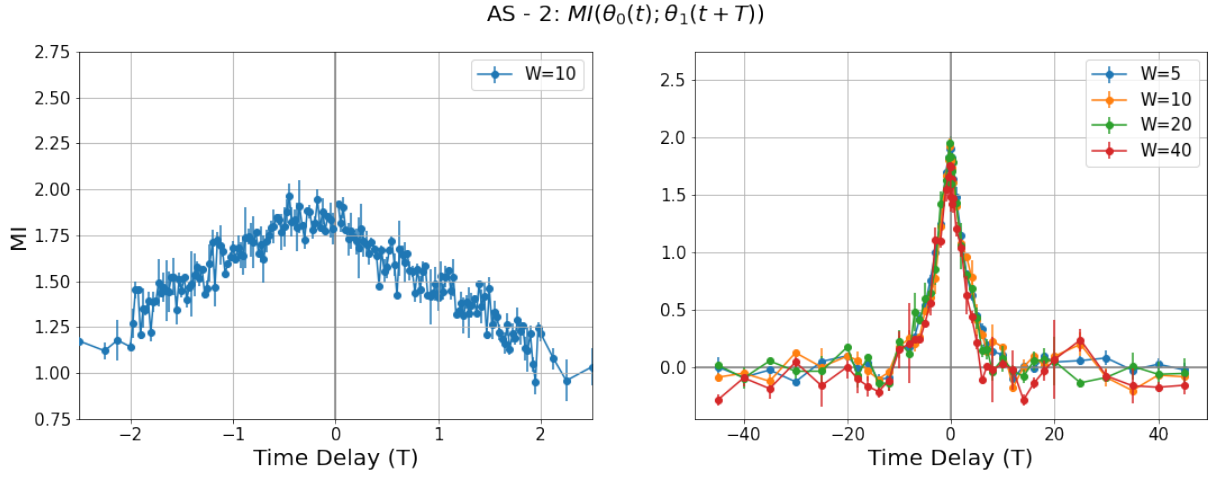


Figure C.11: Mutual information of AS-2.

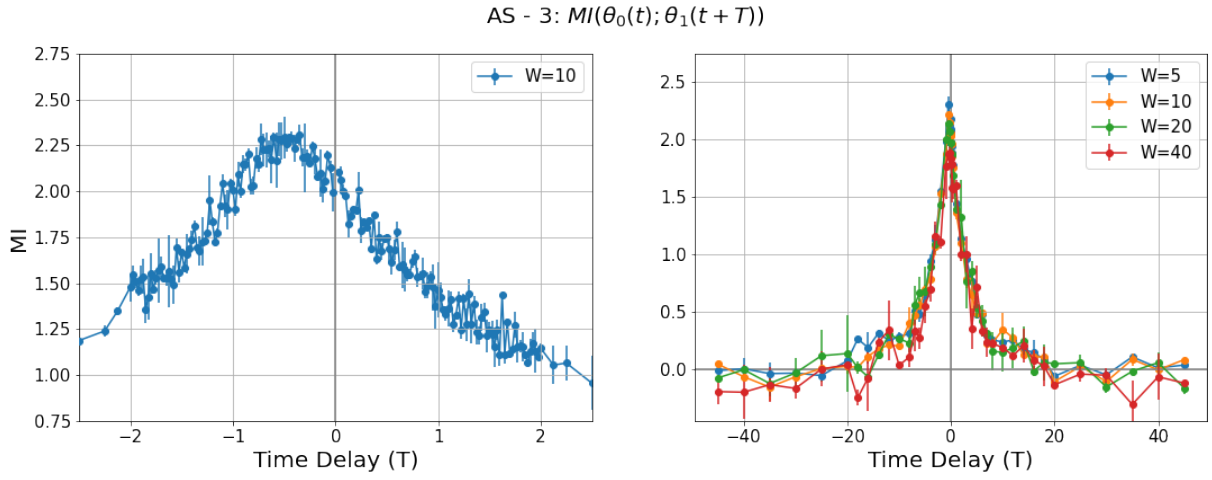


Figure C.12: Mutual information of AS-3.

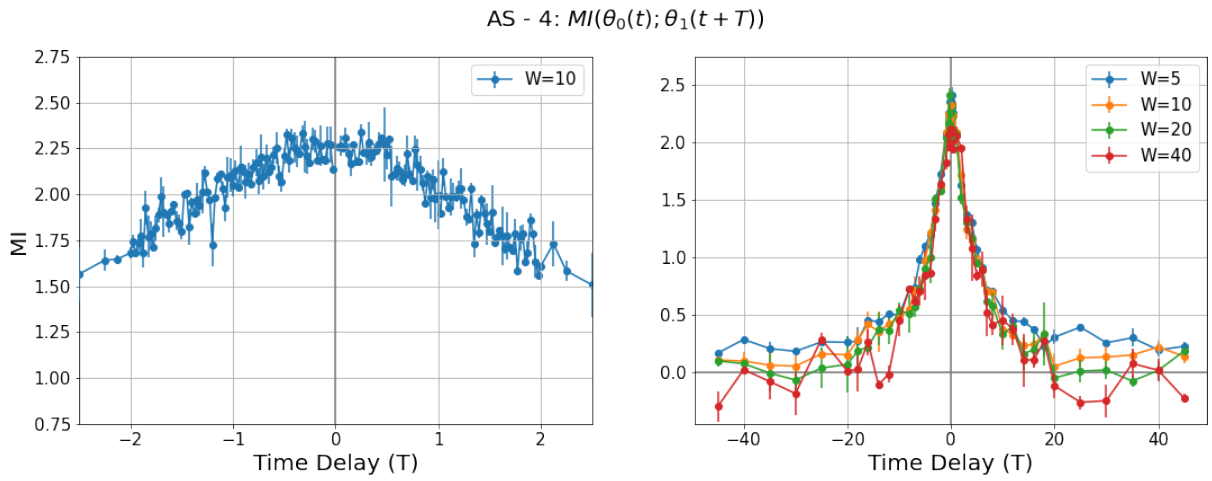


Figure C.13: Mutual information of AS-4.

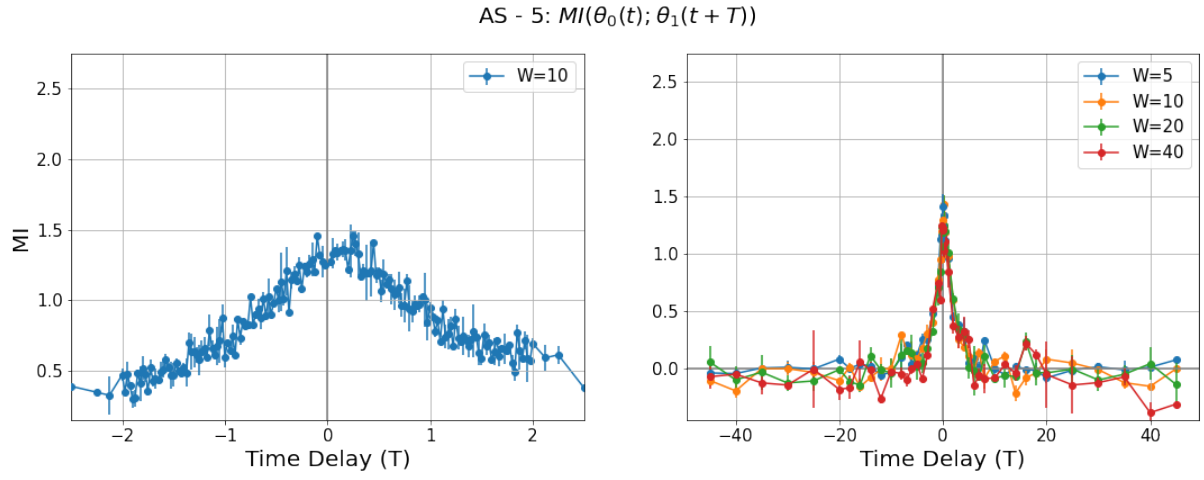


Figure C.14: Mutual information of AS-5. Note different axes in left hand figure compared to other agitated experimental mutual information plots.

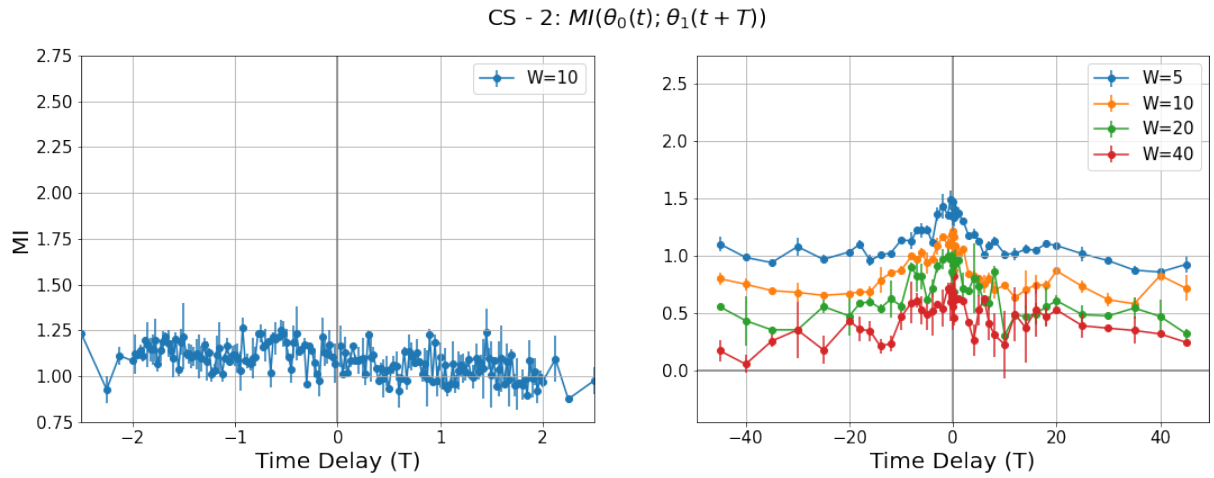


Figure C.15: Mutual information of CS-2.

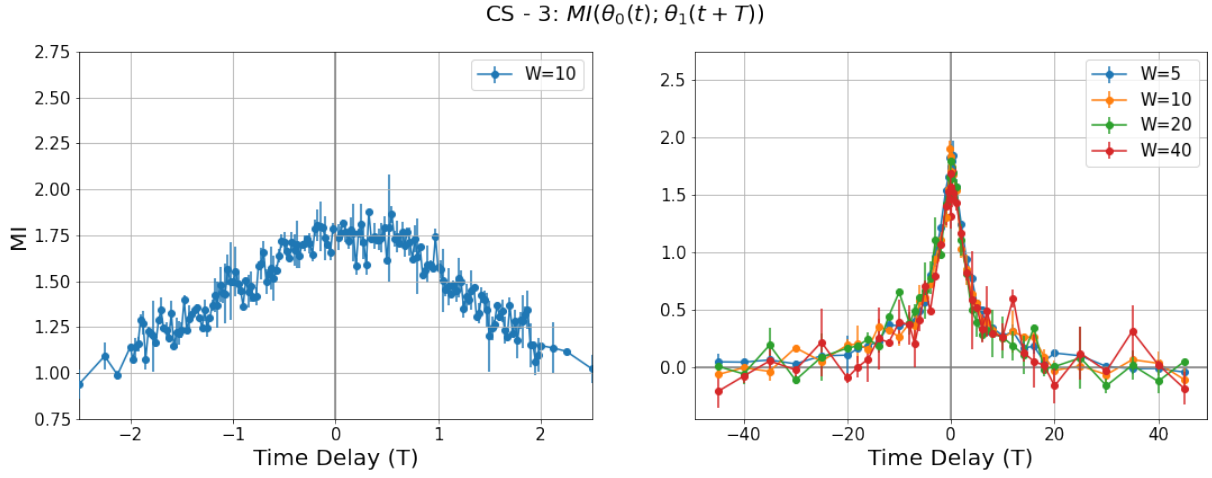


Figure C.16: Mutual information of CS-3.

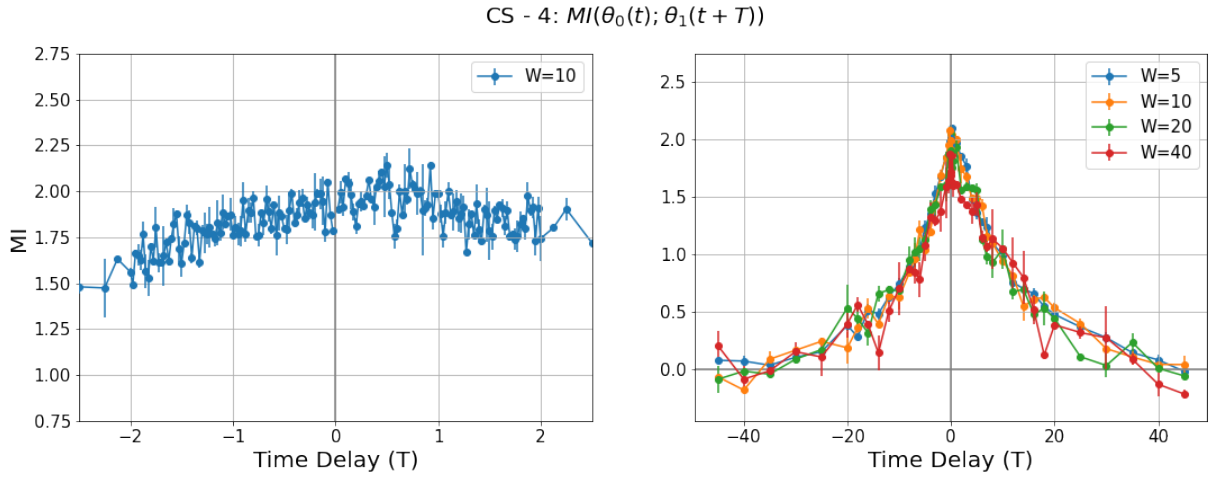


Figure C.17: Mutual information of CS-3.

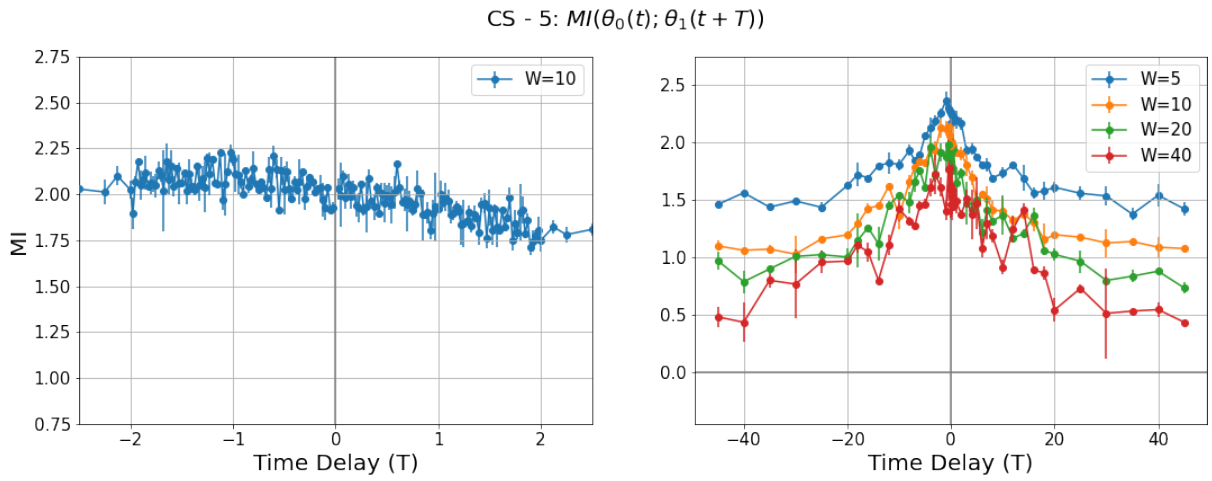
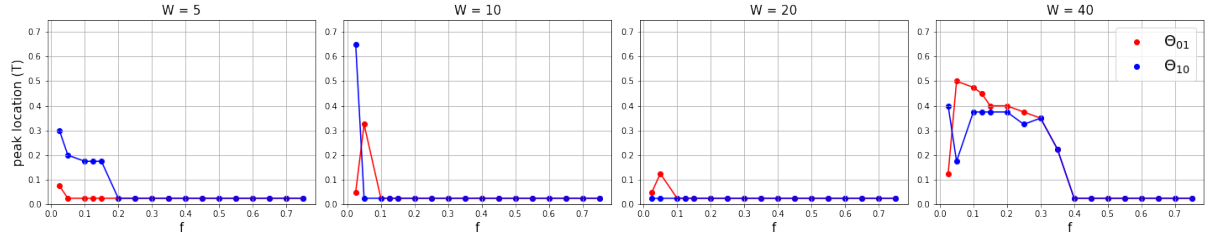
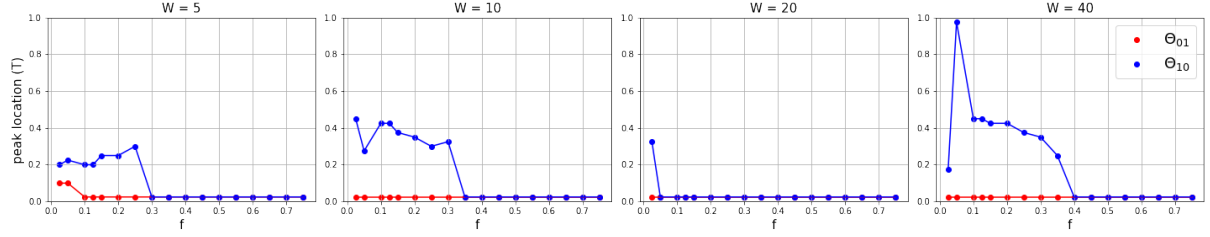


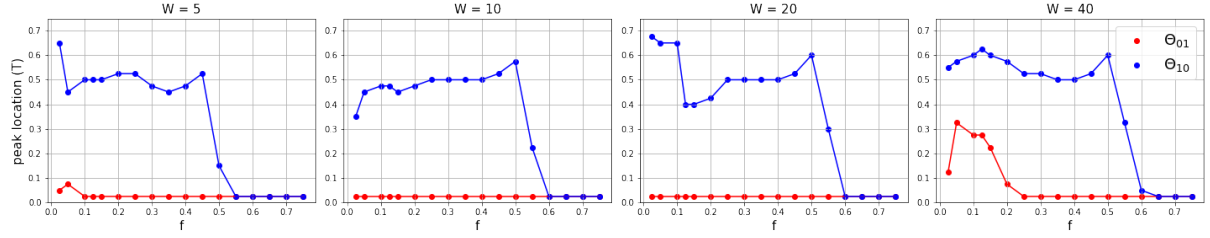
Figure C.18: Mutual information of CS-5.



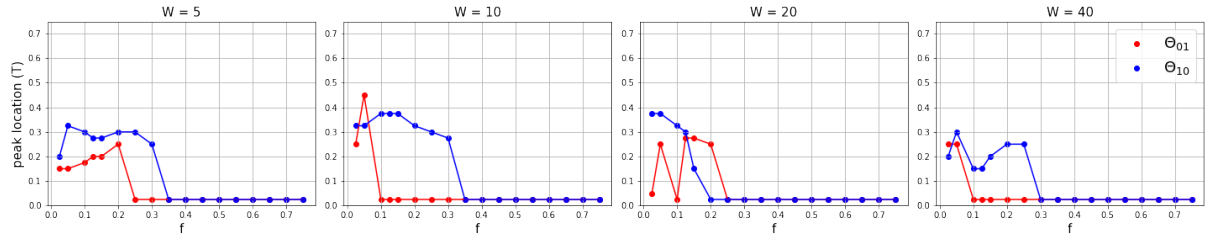
(a) AS-1 LOESS fitted peak locations



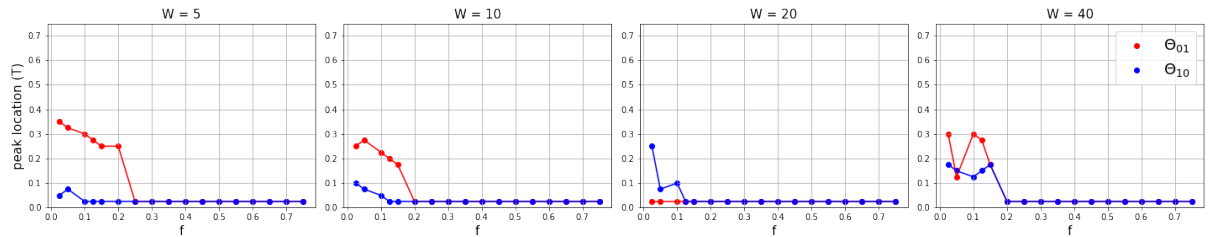
(b) AS-2 LOESS fitted peak locations



(c) AS-3 LOESS fitted peak locations

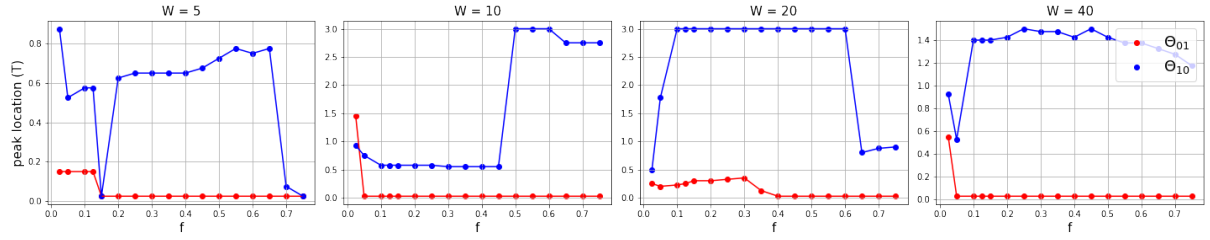


(d) AS-4 LOESS fitted peak locations

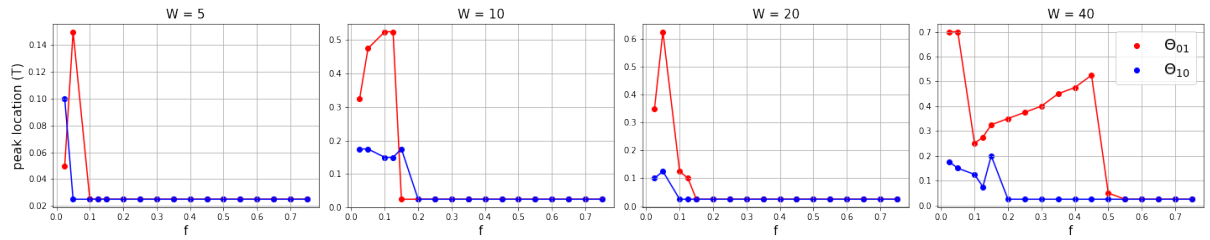


(e) AS-5 LOESS fitted peak locations

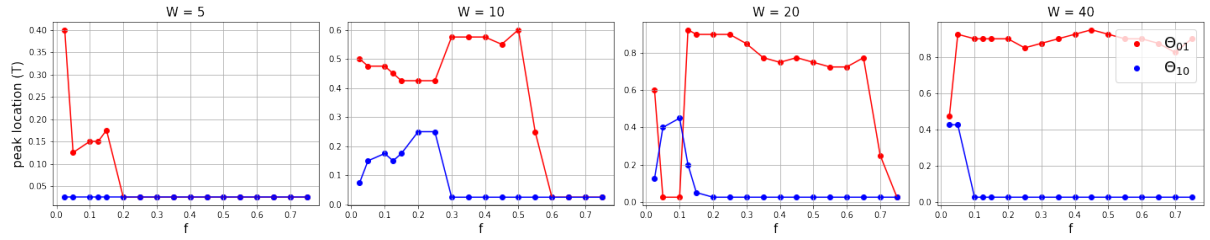
Figure C.19: Varying the data fraction  $f$  changes the location of the maximum mutual information value, which we believe reflects the true signaling timescale ( $T^*$ ) of the experiment. Similarly, choice of separation window  $W$  changes the approximate peak location.



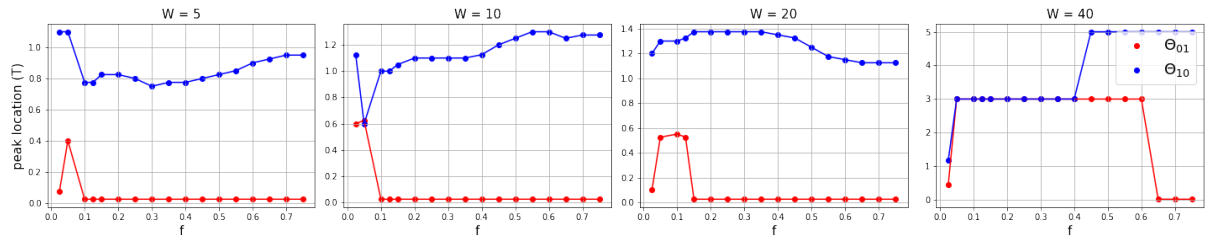
(a) CS-2 LOESS fitted peak locations



(b) CS-3 LOESS fitted peak locations



(c) CS-4 LOESS fitted peak locations



(d) CS-5 LOESS fitted peak locations

Figure C.20: Varying the data fraction  $f$  changes the location of the maximum mutual information value, which we believe reflects the true signaling timescale ( $T^*$ ) of the experiment. Similarly, choice of separation window  $W$  changes the approximate peak location.

## BIBLIOGRAPHY

- [1] J. Adler. Chemotaxis in bacteria. *Science*, 153(3737):708–716, August 1966.
- [2] Z. Alirezaeizanjani, R. Großmann, V. Pfeifer, M. Hintsche, and C. Beta. Chemotaxis strategies of bacteria with multiple run modes. *Science Advances*, 6(22), May 2020.
- [3] C. Anderson and A. Fernandez-Nieves. Social interactions lead to motility-induced phase separation in fire ants. *Nature Communications*, 13(6710).
- [4] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and Z. V. Empirical investigation of starling flocks: a benchmark study in collective animal behaviour. *Animal Behaviour*, 76(1):201–215, 2008.
- [5] A. Berdahl, C. J. Torney, C. C. Ioannou, J. J. Faria, and I. D. Couzin. Emergent sensing of complex environments by mobile animal groups. *Science*, 339(6119):574–576, feb 2013.
- [6] D. D. Biro, J. Sumpter, J. Meade, and T. Guilford. From compromise to leadership in pigeon homing. *Current Biology*, 16:2123–2128.
- [7] N. W. F. Bode, J. J. Faria, D. W. Franks, J. Krause, and A. J. Wood. How perceived threat increases synchronization in collectively moving animal groups. *Proceedings of the Royal Society B: Biological Sciences*, 277(1697):3065–3070, may 2010.
- [8] J. Buck and E. Buck. Mechanism of rhythmic synchronous flashing of fireflies : Fireflies of southeast asia may use anticipatory time-measuring in synchronizing their flashing. *Science*, 159(3821):1319–1327.
- [9] J. Buhl, D. J. T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and S. J. Simpson. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406.
- [10] S. Butail, V. Mwaffo, and M. Porfiri. Model-free information-theoretic approach to infer leadership in pairs of zebrafish. *Physical Review E*, 93(042411).
- [11] I. Buttinoni, G. Volpe, F. Kümmel, G. Volpe, and C. Bechinger. Active brownian motion tunable by light. *Journal of Physics: Condensed Matter*, 24(28):284129, Jun 2012.
- [12] A. Cavagna, A. Cimorelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale. Scale-free correlations in starling flocks. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):11865–11870.
- [13] W.-L. Chen, H. Ko, H.-S. Chuang, H. H. Bau, and D. Raizen. *Caenorhabditis elegans* exhibits positive gravitaxis. *bioRxiv*, 2019.
- [14] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.
- [15] K. Daftari and K. A. Newhall. Self-avoidant memory effects on enhanced diffusion in a stochastic model of environmentally responsive swimming droplets. *Physical Review E*, 105(2).
- [16] A. Dirafzoona, A. Bozkurta, and E. Lobaton. A framework for mapping with biobotic insect networks: From local to global maps. *Robotics and Autonomous Systems*, 88:79 – 96.

- [17] S. J. Ebbens. Active colloids: Progress and challenges towards realising autonomous applications. *Current Opinion in Colloid and Interface Science*, 21:14–23, 2016.
- [18] S. J. Ebbens and J. R. Howse. In pursuit of propulsion at the nanoscale. *Soft Matter*, 6(4):726–738, 2010.
- [19] Y. Fily, S. Henkes, and M. C. Marchetti. Freezing and phase separation of self-propelled disks. *Soft Matter*, 10:2132–2140, 2014.
- [20] Y. Fily and C. M. Marchetti. Athermal phase separation of self-propelled particles with no alignment. *Phys. Rev. Lett.*, 108:235702, Jun 2012.
- [21] D. Geyer, D. Martin, J. Tailleur, and D. Bartolo. Freezing a flock: Motility-induced phase separation in polar active liquids. *Physical Review X*, 9(3).
- [22] F. Ginelli, F. Peruani, M.-H. Pillot, H. Chaté, G. Theraulaz, and R. Bon. Intermittent collective dynamics emerge from conflicting imperatives in sheep herds. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12729–34.
- [23] F. Ginot, I. Theurkauff, D. Levis, C. Ybert, L. Bocquet, L. Berthier, and C. Cottin-Bizonne<sup>1</sup>. Nonequilibrium equation of state in suspensions of active colloids. *Physical Review X*, 5(011004).
- [24] J. G.-J. Godin, L. J. Classon, and M. V. Abrahams. Group vigilance and shoal size in small caracin fish. *Behaviour*, 104(1/2):29–40.
- [25] R. Grima. Strong-coupling dynamics of a multicellular chemotactic system. *Phys. Rev. Lett.*, 95:128103, Sep 2005.
- [26] D. P. Häder. Polarotaxis, gravitaxis and vertical phototaxis in the green flagellate, euglena gracilis. *Archives of Microbiology*, 147(2):179–183, 1987.
- [27] W. Hamilton. Geometry for the selfish herd. *Journal of Theoretical Biology*, 31(2):295–311, may 1971.
- [28] Z. Hao, S. Mayya, G. Notomista, S. Hutchinson, M. Egerstedt, and A. Ansari. Controlling collision-induced aggregations in a swarm of micro bristle robots. *IEEE Transactions on Robotics*, 39(1):590–604, 2023.
- [29] E. Hildebrand and N. Dencher. Two photosystems controlling behavioural responses of halobacterium halobium. *Nature*, 257(5521):46–48, 1975.
- [30] B. V. Hokmabad, M. J. R. Dey, M. A. D. Mohanty, K. A. Baldwin, D. Lohse, and C. C. Maass. Emergence of bimodal motility in active droplets. *arXiv:2005.12721v2*, 2020.
- [31] J. R. Howse, R. A. L. Jones, A. J. Ryan, T. Gough, R. Vafabakhsh, and R. Golestanian. Self-motile colloidal particles: From directed propulsion to random walk. *Physical Review Letters*, 99(4):048102–, 07 2007.
- [32] E. M. Information. A. kraskov and h. stögbauer and p. grassberger. *Physical Review E*, 69(066138).
- [33] A. Izzet, P. G. Moerman, P. Gross, J. Groenewold, A. D. Hollingsworth, J. Bibette, and J. Brujic. Tunable persistent random walk in swimming droplets. *Physical Review X*, 10(2):021035, May 2020.
- [34] A.-Y. Jee, Y.-K. Choa, S. Granicka, and T. Tlusty. Catalytic enzymes are active matter. *PNAS*, 115(46):E10812–E10821, November 2018.

- [35] C. Jin, C. Kruger, and C. C. Maass. Chemotaxis and autochemotaxis of self-propelling droplet swimmers. *Proceedings of the National Academy of Sciences*, 114(20):5089–5094, May 2017.
- [36] H. Ke, S. Ye, R. L. Carroll, and K. Showalter. Motion analysis of self-propelled Pt-silica particles in hydrogen peroxide solutions. *Journal of Physical Chemistry A*, 114(17):5462–5467, April 2010.
- [37] T. Kolb and D. Klotsa. Active binary mixtures of fast and slow hard spheres. *Soft Matter*, 16:1967–1978, 2020.
- [38] W. T. Kranz, A. Gelimson, K. Zhao, G. C. L. Wong, and R. Golestanian. Effective dynamics of microorganisms that interact with their own trail. *Phys. Rev. Lett.*, 117:038101, Jul 2016.
- [39] J. Krause, D. Hoare, S. Krause, C. K. Hemelrijk, and D. I. Rubenstein. Leadership in fish shoals. *Fish and Fisheries*, 1(1):82–89, 2000.
- [40] G. S. R. C. Leblond. Individual leadership and boldness in shoals of golden shiners (*notemigonus crysoleucas*). *Behaviour*, 143(10):1263–1280, 2006.
- [41] B. Liebchen and H. Löwen. Synthetic chemotaxis and collective behavior in active matter synthetic chemotaxis and collective behavior in active matter. *Accounts of Chemical Research*, 51(12):2982–2990, October 2018.
- [42] H. Löwen. Inertial effects of self-propelled particles: From active brownian to active langevin motion. *The Journal of Chemical Physics*, 152(4):040901, January 2020.
- [43] A. E. Magurran and T. J. Pitcher. Provenance, shoal size and the sociobiology of predator-evasion behaviour in minnow shoals. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 229(1257):439–65.
- [44] P. G. Moerman, H. W. Moyses, E. B. van der Wee, D. G. Grier, A. van Blaaderen, W. K. Kegel, J. Groenewold, and J. Bruijic. Solute-mediated interactions between active droplets. *Phys. Rev. E*, 96:032607, Sep 2017.
- [45] V. Mwafo, J. Keshavan, T. L. Hedrick, and S. Humbert. Detecting intermittent switching leadership in coupled dynamical systems. *Nature Scientific Reports*, 8:10338.
- [46] U. of Minnesota Duluth. Increasing golden shiner bait production in minnesota.
- [47] N. Orange and N. Abaid. A transfer entropy analysis of leader-follower interactions in flying bats. *The European Physical Journal Special Topics*, 224:3279–3293.
- [48] C. Orosy-Fildes and R. W. Allan. Psychology of computer use: Xii. videogame play: Human reaction time to visual stimuli. *Perceptual and Motor Skills*, 69:243–247.
- [49] J. Palacci, S. Sacanna, S.-H. Kim, G.-R. Yi, D. J. Pine, and P. M. Chaikin. Light-activated self-propelled colloids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2029), November 2014.
- [50] J. Palacci, S. Sacanna, A. P. Steinberg, D. J. Pine, and P. M. Chaikin. Living crystals of light-activated colloidal surfers. *Science*, 339(6122):936–940, February 2013.
- [51] B. Partridge. The effect of school size on the structure and dynamics of minnow schools. *Animal Behavior*, 28:68–77.

- [52] K. Passino, T. Seeley, and P. Visscher. Swarm cognition in honey bees. *Behavioral Ecology and Sociobiology*, 62:401–414, 2008.
- [53] Patteson, A. E., Gopinath, A., Goulian, M., Arratia, and P. E. Running and tumbling with e. coli in polymeric solutions. *Scientific Reports*, 5(1):15761, 2015.
- [54] W. F. Paxton, K. C. Kistler, C. C. Olmeda, A. Sen, S. K. S. Angelo, Y. Cao, T. E. Mallouk, P. E. Lammert, and V. H. Crespi. Catalytic nanomotors: Autonomous movement of striped nanorods. *Journal of the American Chemical Society*, 126(41):13424–13431, September 2004.
- [55] C. M. Postlethwaite, P. Brown, and T. E. Dennis. A new multi-scale measure for analysing animal movement data. *Journal of Theoretical Biology*, 317:175–185.
- [56] M. S. R, B. E. Jackson, and T. L. Hedrick. The mechanics and behavior of cliff swallows during tandem flights. *Journal of Experimental Biology*, 217:2717–2725.
- [57] G. S. Reebs. Can a minority of informed leaders determine the foraging movements of a fish shoal? *Animal behaviour*, 59(2):403–409, 2000.
- [58] S. Roy, K. Howes, R. Müller, S. Butail, and N. Abaid. Extracting interactions between flying bat pairs using model-free methods. *Entropy*, 21(42).
- [59] D. Saintillan and M. J. Shelley. Instabilities and pattern formation in active particle suspensions: Kinetic theory and continuum simulations. *Physical Review Letters*, 100(178103).
- [60] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.
- [61] A. Sen, M. Ibele, Y. Hong, and D. Velegol. Chemo and phototactic nano/microbots. *Faraday Discussions*, 143(0):15–27, 2009.
- [62] A. Sengupta, S. van Teeffelen, and H. Löwen. Dynamics of a microorganism moving by chemotaxis in its own secretion. *Phys. Rev. E*, 80:031122, Sep 2009.
- [63] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3).
- [64] D. P. Singh, U. Choudhury, P. Fischer, and A. G. Mark. Non-equilibrium assembly of light-activated colloidal mixtures. *Advanced Materials*, 29(32):1701328, June 2017.
- [65] T. Speck, A. M. Menzel, U. Bialk'e, and H. Löwen. Dynamical mean-field theory and weakly non-linear analysis for the phase separation of active brownian particles. *Journal of Chemical Physics*, 142(224109).
- [66] R. L. Stavis and R. Hirschberg. Phototaxis in chlamydomonas reinhardtii. *Journal of Cell Biology*, 59:367–377, 1973.
- [67] J. Stenhammar, A. Tiribocchi, R. J. Allen, D. Marenduzzo, and M. E. Cates. Continuum theory of phase separation kinetics for active brownian particles. *Physical Review Letters*, 111(145702).
- [68] J. Sun and E. Boltt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2000.
- [69] P. A. Svensson, I. Barber, and E. Forsgren. Shoaling behaviour of the two-spotted goby. *Journal of Fish Biology*, 56(6):1477–1487.

- [70] G. Theraulaz. A brief history of stigmergy. *Artificial Life*, 5(2):97–116.
- [71] I. Theurkauff, C. Cottin-Bizonne, J. Palacci, C. Ybert, and L. Bocquet. Dynamic clustering in active colloidal suspensions with chemical signaling. *Physical Review Letters*, 108(268303).
- [72] I. Theurkauff, C. Cottin-Bizonne, J. Palacci, C. Ybert, and L. Bocquet. Dynamic clustering in active colloidal suspensions with chemical signaling. *Physical Review Letters*, 108(26):268303, June 2012.
- [73] D. D. Thomas and A. P. Peterson. Chemotactic auto-aggregation in the water mould achlya. *Microbiology*, 136(5):847–853, May 1990.
- [74] S. Thutupalli, R. Seemann, and S. Herminghaus. Swarming behavior of simple model squirmers. *New Journal of Physics*, 13(7):073021, July 2011.
- [75] Y. Tsori and P.-G. de Gennes. Self-trapping of a single bacterium in its own chemoattractant. *Europhysics Letters (EPL)*, 66(4):599–602, 2004.
- [76] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(1226).
- [77] T. Vicsek and A. Zafeiris. Collective motion. *Physics Reports*, 517(3-4):71–140.
- [78] A. J. W. Ward, D. J. T. Sumpter, I. D. Couzin, P. J. B. Hart, and J. Krause. Quorum decision-making facilitates information transfer in fish shoals. *Proceedings of the National Academy of Sciences*, 105(19):6948–6953, may 2008.
- [79] Y. Yang, V. Lam, M. Adomako, R. Simkovsky, A. Jakob, N. C. Rockwell, S. E. Cohen, A. Taton, J. Wang, J. C. Lagarias, A. Wilde, D. R. Nobles, J. J. Brand, and S. S. Golden. Phototaxis in a wild isolate of the cyanobacterium *synechococcus elongatus*. *PNAS*, 115(52):E12378–E12387, December 2018.
- [80] J. Zhang, E. Luijten, B. A. Grzybowski, B. A., , and S. Granick. Active colloids with collective mobility status and research opportunities. *Chem. Soc. Rev*, 46(18):5551–5569, 2017.
- [81] F. Zhao, W. Rong, L. Wang, and L. Sun. Magnetic actuated shape-memory helical microswimmers with programmable recovery behaviors. *Journal of Bionic Engineering*, 18:799–811.