

**PHYSICS-INSPIRED MACHINE LEARNING OF PARTIAL DIFFERENTIAL  
EQUATIONS**

A Dissertation  
Presented to  
The Academic Faculty

By

Matthew Golden

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
School of Physics

Georgia Institute of Technology

August 2023

© Matthew Golden 2023

**PHYSICS-INSPIRED MACHINE LEARNING OF PARTIAL DIFFERENTIAL  
EQUATIONS**

Thesis committee:

Dr. Roman Grigoriev  
School of Physics  
*Georgia Institute of Technology*

Dr. Elisabetta Matsumoto  
School of Physics  
*Georgia Institute of Technology*

Dr. Michael Schatz  
School of Physics  
*Georgia Institute of Technology*

Dr. Alberto Fernandez-Nieves  
Department of Physics  
*University of Barcelona*

Dr. Kurt Wiesenfeld  
School of Physics  
*Georgia Institute of Technology*

Date approved: July 14, 2023

## ACKNOWLEDGMENTS

I must acknowledge my family. My grandparents Charles and Janet Peterson supported me financially in life and death and showed unwavering support of my academic journey. My brothers Bradford and Alan were always willing to entertain my mathematical ramblings, and my parents Michael and Jayna allowed me to not be an engineer. I am especially grateful to Molly for years of personal support in all forms. My friends have made Atlanta a second home. Any list I attempt to give will be incomplete.

I thank my academic advisors Roman Grigoriev and Michael Schatz for providing encouragement and direction in my research. The progress I made would have been impossible without their frequent feedback and guidance. I thank my long-time colleagues Christopher Crowley, Joshua Pughe-Sanford, Dmitriy Zhigunov, and Mateo Reynoso for countless fruitful discussions of our research. I thank Daniel Gurevich for our weekly discussions on model discovery. Lastly, I thank Jyothishraj Nambisan for his experimental contributions and illuminating conversations.

## TABLE OF CONTENTS

<b>Acknowledgments</b> . . . . .	iii
<b>List of Tables</b> . . . . .	vi
<b>List of Figures</b> . . . . .	viii
<b>Summary</b> . . . . .	xii
<b>Chapter 1: Introduction</b> . . . . .	1
<b>Chapter 2: Symmetry-Covariant Tensor Libraries</b> . . . . .	7
2.1 Justification . . . . .	7
2.2 Index Notation . . . . .	9
2.3 The Diffeomorphism Group and $GL(n, \mathbb{R})$ . . . . .	16
2.4 $O(n)$ . . . . .	20
2.5 $SO(n)$ . . . . .	23
2.6 The Homogeneous Galilean Group . . . . .	27
2.7 $SU(2)$ and $SL(2, \mathbb{R})$ spinors . . . . .	29
<b>Chapter 3: Feature Matrix Construction and Sparse Regression</b> . . . . .	35
3.1 Matrix Construction . . . . .	35
3.2 Introduction to Regression . . . . .	38

3.3	Basis Pursuit Denoising and LASSO . . . . .	40
3.4	Sequentially Thresholded Least Squares . . . . .	42
3.5	Greedy Iterative SVD Regression . . . . .	44
3.6	Reverse Regression: a holistic approach to sparse regression . . . . .	47
3.7	Hyperparameter Selection . . . . .	52
3.8	Data-Dependent Weight Functions . . . . .	56
<b>Chapter 4: Applications of SPIDER . . . . .</b>		<b>60</b>
4.1	Direct Numerical Simulations of Turbulence . . . . .	60
4.2	Active Nematics . . . . .	64
<b>Chapter 5: Conclusion . . . . .</b>		<b>86</b>
<b>Appendices . . . . .</b>		<b>89</b>
	Appendix A: SPIDER Code . . . . .	90
<b>References . . . . .</b>		<b>92</b>

## LIST OF TABLES

2.1	<b>Low-dimensional representations of the diffeomorphism group of a general manifold.</b> Without an invertible quadratic form, there is no natural bijection from covariant to contravariant spaces. For a total rank less than 2, there are 8 distinct irreducible representations physics could live in. . . . .	18
2.2	<b>The irreducible tensors up to rank 2 of the <math>O(n)</math> symmetry group.</b> The rank-2 tensors split into three irreducible tensor spaces, however the representation proportional to the metric is redundantly isomorphic to the scalar library. . . . .	21
2.3	<b>The irreducible tensors up to rank 2 of the <math>SO(2) \cong U(1)</math> symmetry group.</b> It is an interesting feature of this group that only one non-redundant representation exists for each rank. Tensor multiplication for these 1D $\mathbb{C}$ spaces is just ordinary complex multiplication. . . . .	25
2.4	<b>The irreducible tensors up to rank 2 of the <math>SO(3)</math> symmetry group.</b> Like $SO(2)$ , there is a single representation per rank warranting investigation. . . . .	26
2.5	<b>The irreducible tensors up to rank 2 of the <math>SU(2)</math> symmetry group.</b> . . . . .	33
2.6	<b>The irreducible tensors up to rank 2 of the <math>SL(2, \mathbb{C})</math> symmetry group describing relativistic spinors.</b> . . . . .	34
3.1	<b>Characteristics of some exact solutions to BPDN.</b> $\lambda$ is the regularization parameter, $\ \mathbf{c}\ _0$ is the number of nonzero terms in $\mathbf{c}$ , $r$ is the relative residual, and $\mathbf{c}^+$ is the coefficient vector obtained from the least-squares problem in the same subspace as $\mathbf{c}$ . . . . .	41
4.1	<b>Table of identified coefficients.</b> The uncertainties $s_n$ are estimated by the standard deviation of coefficients when regression is done 100 times with different integration subdomains. The magnitude of terms $\chi_n = \ c_n \mathbf{g}_n\  / \max\{\ c_m \mathbf{g}_m\ \}$ determines how much each term contributes to the relation. . . . .	62

4.2 **The summary of the model libraries and their symmetry properties.**  
The nematic symmetry doubles the number of representations worth studying, although two are neglected in this work. . . . . 66

## LIST OF FIGURES

1.1	<b>Periodic numerical solutions to the three body problem.</b> The bodies have equal masses and units such that $m =  H  = G = 1$ . These solutions were converged with Newton-Krylov iteration until the equations of motion were satisfied in spectral form to machine precision. . . . .	2
1.2	<b>A schematic of the SPIDER algorithm from [37].</b> Tree tensors are constructed from tensor products of tensors and their gradients. All possible contractions of tree tensors are taken to produce prelibraries. Symmetry constructed projection operators project prelibraries into irreducible representations, such that there is one library per irreducible representation. These are then evaluated in the weak form to construct coefficient equations $Gc = 0$ , which are then solved by a sparse regression algorithm to produce parsimonious empirical relations. . . . .	6
3.1	<b>Polynomial envelopes for representative choices of <math>n</math>.</b> The higher the value of $n$ , the fewer grid points are effectively used in averaging. . . . .	36
3.2	<b>STLS applied to the rank-2 library of active nematic data. (a) shows the number of nonzero terms in <math>c</math> as a function of <math>\lambda</math>. (b) shows the relative residual <math>r(c)</math> as a function of <math>\lambda</math>.</b> . . . . .	43
3.3	<b>A schematic of GISR.</b> Terms are greedily removed from the library to obtain a sparse approximate null vector. The SVD is recomputed after every modification of the library. . . . .	45
3.4	<b>Residual curves for two subsequent runs of GISR applied to a rank-2 library.</b> The first run finds a five term relation. After the largest term is discarded from the library, a subsequent run finds a 2-term relation. . . . .	47
3.5	<b>Scaling of the GISR computation time with library size.</b> The dashed line shows a polynomial fit with scaling $n^4$ . . . . .	48
3.6	<b>Final state after integration of the lambda-omega reaction diffusion equations.</b> (a) shows $u$ and (b) shows $v$ . . . . .	49

3.7	<b>Reverse regression curves of the reaction-diffusion library starting from 2-term dominant balances.</b> The red marker indicates where no more terms need added. The blue marker indicates where no more terms can be removed. (a), (b), and (c) start RR with Equation 3.20, Equation 3.22, and Equation 3.24, respectively. . . . .	50
3.8	<b>The Lorenz attractor.</b> (a) A 2D projection of a numerical trajectory generated with RK4 with $dt = 1/200$ . The initial transient is removed so that the trajectory lies very close to the attractor. (b) Numerical evidence that $T = \beta^{-1}$ is a reasonable timescale for $z$ dynamics on the attractor. . . . .	52
3.9	<b>Probability of discovering the correct dynamics of <math>z</math> with various nondimensionalization schemes, noise levels, and number of domains.</b> Integrals for the weak form are computed with 256 points in time and envelope power 6. This corresponds to a nondimensional window size of $\beta t_{\Omega} \approx 3.4$ . Since the library contains 10 terms, each matrix $G$ is of size $d \times 10$ . To estimate the correct discovery probability, subdomains are chosen randomly 16 times and GISR is performed independently for each sampling with $\gamma = 1.5$ . (a) OND. (b) UND. (c) SND. (d) PND. . . . .	55
3.10	<b>Scaling of the relative residual of the trapezoid rule.</b> The points on the nonuniform grid have only been moved up by to 1% of the grid spacing. . . . .	56
3.11	<b>A mask and its derivative computed with the MATLAB function <code>generate_mask</code>.</b> Two points at $x \approx 0.3$ and $x \approx 0.7$ have been flagged as ‘bad’ so that $\psi = 0$ exactly there. (a) shows the constructed mask and (b) shows the derivative estimated with a centered difference. Higher derivatives should be evaluated with regularized difference schemes given the fast variation of $\psi$ . . . . .	57
3.12	<b>A mask constructed to mask topological defects.</b> (a) shows a snapshot of $Q_{11}$ , a component of the nematic tensor $Q_{ij}$ . The color bar is arbitrary. Topological defects can be identified as points with abrupt change in $Q_{11}$ . (b) shows a computed $\psi$ that goes to zero smoothly around these point-like defects. . . . .	59
4.1	<b>Residual curves for the identified relations.</b> Panel(a) shows the residual curve for the energy equation. Panel (b) shows the residual curve of the pressure Poisson equation. Panel (c) shows the residual curve of the phenomenological pressure transport relation. Panel (d) shows the residual curve of the Navier-Stokes equation. . . . .	63

4.2	<b>Experimental snapshot and observable fields.</b>	(a) An experimental image of the microtubules. The complete image is shown, with the blue box highlighting a $-1/2$ topological defect. Panel (b) shows a zoomed-in $-1/2$ defect and the extracted nematic field around it. The extracted director field $\mathbf{n}$ (blue arrows) is clearly unreliable near topological defects. (c) Director field $\mathbf{n}$ (black arrows) and the mask $\psi$ (color) used to remove near-defect data. Panel (d) shows the flow field $\mathbf{u}$ (black arrows) and the corresponding vorticity $\omega = (\nabla \times \mathbf{u})_z$ (color). . . . .	67
4.3	<b>Remedying the branch cuts of the nematic field for weak-form evaluation.</b>	Panel (a) shows the splay scalar $s$ around a topological defect and a visible branch cut. Panel (b) shows $sn_x$ , which is nematic invariant and no longer suffers from a line discontinuity. The colorbar is arbitrary. . . . .	72
4.4	<b>The relative residual curve as a function of the number of terms in the relation.</b>	The plots corresponding to discovered equations are (a) (Equation 4.14), (b) (Equation 4.15), (c) (Equation 4.17), (d) (Equation 4.19), (e) (Equation 4.21), (f) (Equation 4.22), (g) (Equation 4.25), (h) (Equation 4.26), and (i) (Equation 4.29). The red cross indicates the corresponding parsimonious relation. . . . .	77
4.5	<b>Correlation between the predicted nematic field and the observed nematic field.</b>	$Q(t)$ is the experimentally observed field and $\tilde{Q}(t)$ is numerically integrated via the leapfrog method from the first two frames using the state velocity predicted by Equation 4.19 using the experimental flow field. $\langle f g \rangle$ is the normalized inner product in the sense of Dirac. The lowest correlation between numerical predictions and experiment is 0.93. This minimum lowers to 0.88 if the constant extension constraint is approximately solved using experimental boundary conditions for the streamfunction $\psi$ . Note that good correlation is found even after the correlation with the initial state has completely vanished. . . . .	79
4.6	<b>Histogram of flow gradient scalars.</b>	The boundaries in space and time have been removed, as well as regions with $\psi < 0.2$ . The scale of the PDF is arbitrary but consistent. . . . .	82
4.7	<b>Predicted flow around confined topological defects.</b>	(a) An ideal $+1/2$ defect generates two counter-rotating vortices responsible for self-driven motion. (b) An ideal $-1/2$ defect generates a hexagonal flow. (c) An experimental flow field in the comoving, corotating frame of a $+1/2$ defect. Two counter-rotating vortices are visible. (d) An experimental flow field in the comoving, corotating frame of a $-1/2$ defect. Alternating vortices are visible, although perfect hexagonal symmetry is not realized. The velocity is shown in the comoving, corotating frame since in such a frame $A_{ij}$ locally dominates the flow structure. . . . .	84

4.8 **Verification of relations in strong form.** Panel (a) shows the experimental frame of interest. Panel (b) shows the strong form of  $\nabla_i u_i$ , which has fluctuations over small length scales but no recognizable structure away from  $+1/2$  defects. Panel (c) shows the time derivative  $\partial_t \theta = \mathbf{n} \times \partial_t \mathbf{n}$  computed by finite differencing the director field in time, and Panel (d) shows that same quantity, but calculated with spatial finite differencing from Equation 4.19. Panels (e) and (f) show  $c_1^{(1)} Q_{xx}$  and  $\bar{A}_{kl} \bar{Q}_{kl} Q_{xx}$  of Equation 4.25. Panels (g) and (h) show the corresponding  $xy$  components. . . . . 85

## SUMMARY

This dissertation discusses the Sparse Physics-Informed Discovery of Empirical Relations (SPIDER) algorithm, which is a technique for data-driven discovery of governing equations of physical systems. SPIDER combines knowledge of symmetries, physical constraints like locality, the weak formulation of differential equations, and sparse regression to construct mathematical models of spatially extended physical systems. SPIDER is a valuable tool in synthesizing scientific knowledge as demonstrated by its applications.

Data-driven model discovery is a three-step process. First, libraries of terms are constructed using available physical fields. The symmetries of a system allow libraries to be projected into independently transforming spaces, known as irreducible representations. This breaks relations down into their indivisible parts; each minimal physical relation is learned independently to reduce implicit bias. A library of nonlinear functions is constructed for each irreducible representation of interest.

Second, each library term is evaluated in the weak formulation. SPIDER is aimed at experimental systems with inherently noisy data making accurate estimation of derivatives difficult. The weak formulation solves this problem: library terms are integrated over spacetime domains with flexible weight functions. Integration by parts can avoid numerical differentiation in many situations and increases robustness to noise by orders of magnitude. Clever weight functions can remove discontinuities and even entirely remove unobserved fields from analysis. Third, a sparse regression algorithm can find parsimonious relations ranging from dominant balances to multi-scale quantitatively accurate relations.

Applications to direct numerical simulation of 3D fluid turbulence and experimental 2D active nematic turbulence are presented. SPIDER recovered complete mathematical models of both systems. The active nematic system is of particular interest; SPIDER identified a 2D description contradicting widely accepted theoretical descriptions used for over a decade. SPIDER facilitated the discovery of a new physical constraint on the fluid flow.

# CHAPTER 1

## INTRODUCTION

The natural world cannot help but be described by mathematics. Kepler's breakthrough work on planetary motion in the early 1600's heralded a revolution in mathematical modeling the physical world. The development of calculus allowed the description of continuous change in space and time through the derivative operator. Newton showed that basic physics can be described by relations between derivatives, known as differential equations. Differential equations lie at the foundation of almost every physical theory. Various differential equations capture chemical reactions [1], population dynamics [2], fluid dynamics [3], cosmology [4], the spread of epidemics [5], and countless other phenomena. Differential equations are particularly apt at describing change in time, and therefore predicting the future.

Obtaining a differential equation describing a real physical system is no small feat. There is always some approximation being made and some small effects being ignored. A model is first made that captures the dominant processes. The typical evolution of this mathematical model is then via a feedback loop. Observations are made and compared to current predictions. If the predictions agree, more data is taken. If the predictions do not agree, then the model is updated to account for new processes until there is agreement between experiment and theory.

A primary difficulty of this approach is many nonlinear differential equations do not have general analytic solutions; a computer is required to find approximate solutions via direct numerical integration. Perhaps the most well-known example of such a system is the three body problem of celestial mechanics investigated by Poincaré [6]. Some short closed orbits of the three-body problem can be seen in Figure 1.1. Making predictions for high-dimensional systems, in particular fluid dynamics and numerical astrophysics, can be

incredibly expensive. Full 3D simulations of black hole binaries can take days to a month to complete [7]. Further, models usually depend on a number of physical parameters such as viscosity, heat capacity, or density. Model predictions can be extremely sensitive to the values of these parameters, such that small changes completely change the qualitative behaviour of solutions via bifurcations [8, 9].

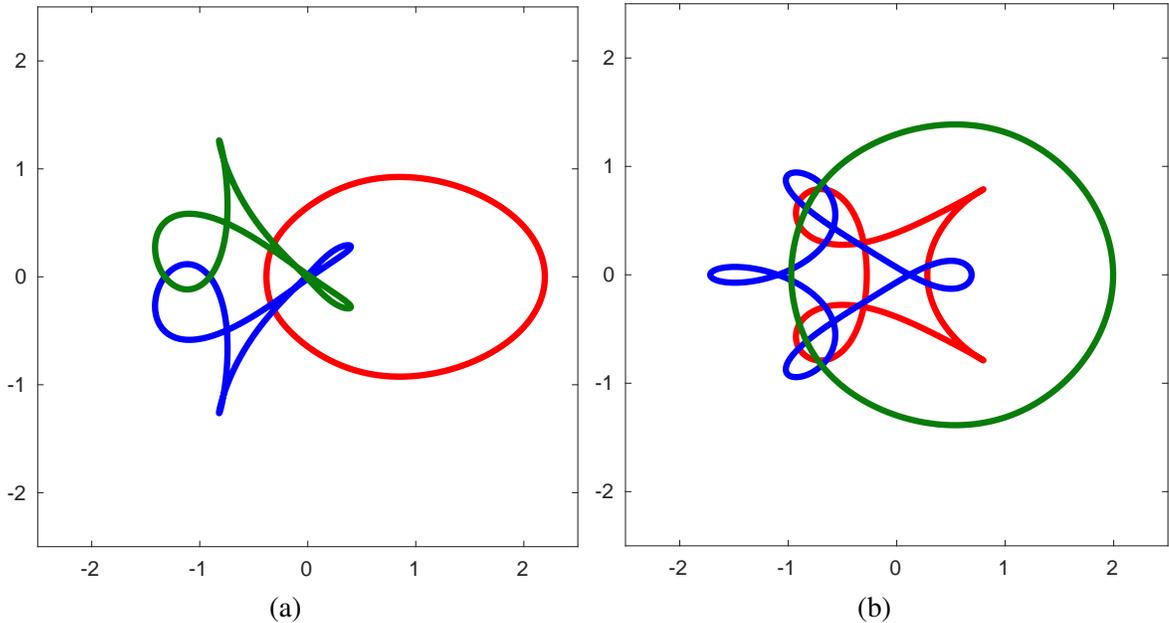


Figure 1.1: **Periodic numerical solutions to the three body problem.** The bodies have equal masses and units such that  $m = |H| = G = 1$ . These solutions were converged with Newton-Krylov iteration until the equations of motion were satisfied in spectral form to machine precision.

There has been a paradigm shift in recent years to develop methods that take the humans out of the feedback loop. The process is short-circuited to find qualitatively and quantitatively accurate models directly from the experimental data. Numerous approaches have now been developed to infer equations, which vary in sensitivity to noise, computation time, constraint-enforcement, and explored equation space.

Perhaps the most common data-driven approach to modeling is to train a neural network to fit some physical measurement. Examples include fitting the Reynolds stress anisotropy in Reynolds-averaged turbulence simulations [10], learning closures to projections of the Boltzmann equation [11], extracting critical exponents by learning the renormalization

group [12], or forecasting system dynamics with physics-informed neural networks [13]. In recent years, emphasis has been placed on neural networks exactly preserving symmetries [14, 10, 11]. While neural networks can be trained efficiently and fit the data convincingly, they lack interpretability. There is not a clear means of communicating new physics learned by a neural network, although effective network weight sparsification is a natural step towards this [15, 16]. One path to interpretability is symbolic regression via genetic algorithms. Early work showed that genetic algorithms can learn the dynamics, Hamiltonian, and Lagrangian equations for low-dimensional systems without imposing any physics knowledge [17, 18]. Recent genetic programming has learned 100 relations from the *Feynman lectures on physics* from Compton scattering to the Friedman equation [19].

The genetic algorithms can find arbitrarily complex nonlinear relations by modeling relations as trees. However, genetic programming struggles to find equations that are sums of more than 2 complex terms because of complexity considerations; such relations will be found very late into a search. Many relations are more straightforwardly represented as a linear combinations of nonlinear terms:  $c_1 \mathbf{f}_1 + c_2 \mathbf{f}_2 + \dots + c_n \mathbf{f}_n = 0$ . The automated determination of  $\mathbf{f}_n$  and values of coefficients  $c_n$  has been of significant interest [20, 17, 21, 22]. A common problem is the determination of the state dynamics  $\partial_t z_i$ . The time derivatives are fit to a linear combination of ad hoc candidate functions  $\{z_1, z_2, \ln(1 + |z_3|), \dots\}$  specified by the modeler. The library terms and the time derivative are evaluated at various points in time to create an over-determined system of equations.

$$\mathbf{g} = \begin{bmatrix} | \\ \partial_t z_1 \\ | \end{bmatrix} = \begin{bmatrix} | & | & | & \dots \\ z_1 & z_2 & \ln(1 + |z_3|) & \dots \\ | & | & | & \dots \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{bmatrix} = G\mathbf{c} \quad (1.1)$$

Perhaps the most common algorithm for coefficient determination is the Sparse Identification of Nonlinear Dynamics (SINDy) [23]. Solutions with sparse  $\mathbf{c}$  are found by sequential

thresholding, where the system is iteratively solved numerically and small values of  $c$  are set to zero. While simplistic, SINDy showed the power of data-driven sparse regression for rapidly and robustly deducing evolution equations in the form of ODEs, especially for low-dimensional systems [24]. Many modifications have been proposed in the literature. An incomplete list is Weak-SINDy for learning the weak-formulation of differential equations [25, 26, 27], SINDy-PI for learning implicit and rational dynamics [28], and Ensemble-SINDy for learning in the low data limit [29]. SINDy is ultimately a linear regression algorithm with no knowledge of physics. Its deficiencies are its susceptibility to noise and lack of guidance on library construction: coefficient determination will only work if the correct terms are present in the library. This is a generic feature of regression, but it can be helped by a systematic library construction.

This dissertation explores the symbolic regression method Sparse Physics-Informed Discovery of Empirical Relations (SPIDER), which uses physics domain knowledge of locality and symmetry to constrain the model discovery process. This builds on a body of machine learning literature devoted to spatially extended nonequilibrium systems [30, 31, 32, 33, 34, 35, 36]. The symmetries of the physical system are combined with the experimental measurements to identify a complete set of interpretable quantitatively accurate physical relations including evolution equations and spatial constraints. SPIDER is a *fundamentally physics-constrained* scheme for the identification of partial differential equations and algebraic relations. Figure 1.2 provides a visual map to the SPIDER algorithm. SPIDER has several distinguishing characteristics compared to other model discovery algorithms. It is the combination of these that gives SPIDER its versatility and power.

- Many model discovery algorithms fit a single term  $g$ , usually a time derivative, to a evaluations of a library of terms  $G$ . This coefficient problem encourages unwarranted assumptions about governing equations. For example, spatial constraints cannot be learned if they involve no time derivative. SPIDER seeks to instead solve the homogeneous problem  $Gc = 0$ ; all library terms are put on equal footing. This is partially

solved by SINDy-PI where each library term is effectively independently fit via the inhomogeneous problem [28], but this is computationally expensive.

- Many model discovery algorithms usually consider a single library [35, 36]. For a collection of fields containing scalars, vectors, and higher rank tensors, each tensor space can be considered independently and contains independent physics. An ensemble of libraries can be constructed where each irreducible representation of system symmetries (up to some cutoff) has a symbolic library. Assumptions of symmetry can be relaxed if needed, but even general symmetry assumptions are useful in reducing library size.
- SPIDER explicitly uses the weak formulation of differential equations, in which library terms are integrated against weight functions with compact support. The weak formulation of differential equations has been shown to significantly increase noise robustness [25, 26].

These three principles of homogeneous regression, irreducible representations, and weak evaluation of library terms are the pillars of SPIDER. This manuscript is organized in the following way. Chapter 2 discusses tensor libraries and some low rank irreducible representations of common symmetry groups. Chapter 3 will discuss the weak formulation and sparse regression techniques that have been used with SPIDER. Chapter 4 will discuss specific results from applying SPIDER to numerical 3D turbulence and a 2D experimental active nematic suspension.

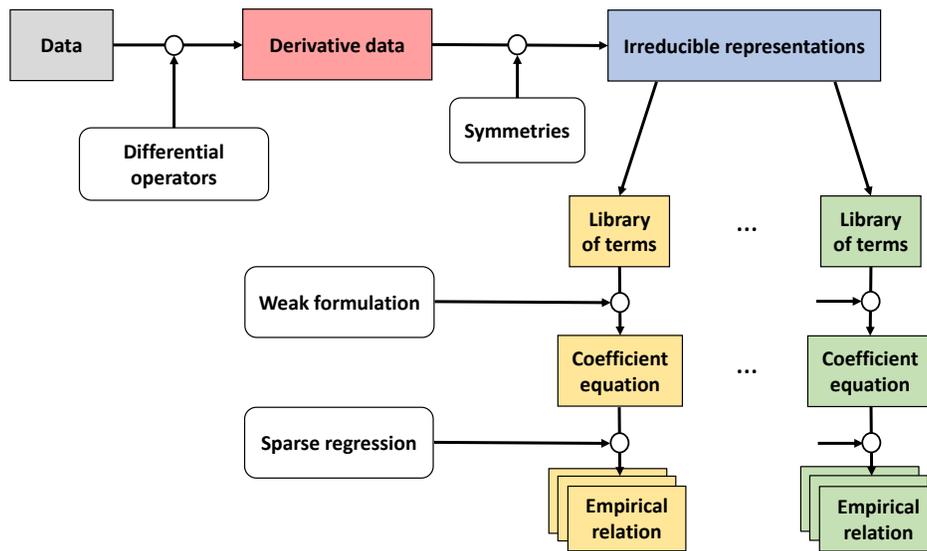


Figure 1.2: **A schematic of the SPIDER algorithm from [37].** Tree tensors are constructed from tensor products of tensors and their gradients. All possible contractions of tree tensors are taken to produce prelibraries. Symmetry constructed projection operators project prelibraries into irreducible representations, such that there is one library per irreducible representation. These are then evaluated in the weak form to construct coefficient equations  $Gc = 0$ , which are then solved by a sparse regression algorithm to produce parsimonious empirical relations.

## CHAPTER 2

### SYMMETRY-COVARIANT TENSOR LIBRARIES

#### 2.1 Justification

The general goal of model discovery via sparse regression is to search through a library of functions of observables  $\mathcal{L} = \{\mathbf{F}_n\}$  and find a sparse vector  $c_n$  such that

$$\sum_n c_n \mathbf{F}_n \approx 0. \quad (2.1)$$

By finding interpretable relations of this form, machine learning accelerates human learning. This task becomes more difficult as the size of library  $\mathcal{L}$  grows: more comparisons must be performed and more data is needed to meaningfully distinguish correlated terms. Symmetry considerations can help fight this scaling of complexity. Suppose a particle in 3D moves in a spherically symmetric potential such that the underlying dynamics take the form

$$\ddot{\mathbf{x}} = -\nabla V(r). \quad (2.2)$$

Attempting to fit the components of the right hand side in Equation 2.2 naively with polynomials up to cubic order in coordinates gives a system of equations with 54 unknowns.

$$\begin{aligned} \ddot{x} &= c_1 + c_2x + c_3y + c_4z + c_5x^2 + c_6xy + c_7y^2 + c_8yz + \dots \\ \ddot{y} &= c_{19} + c_{20}x + c_{21}y + c_{22}z + c_{23}x^2 + c_{24}xy + c_{25}y^2 + c_{26}yz + \dots \\ \ddot{z} &= c_{37} + c_{38}x + c_{39}y + c_{40}z + c_{41}x^2 + c_{42}xy + c_{43}y^2 + c_{44}yz + \dots \end{aligned} \quad (2.3)$$

This is a linear system of equations for  $c_i$ , which can be numerically solved once a set of observations is made. Each component can be fit independently to solve for 18 coefficients

at a time. If noise is present in the measurements, it is unlikely that any of the  $c_n$  will naturally be zero from the least-squares solution. Note however that most of the terms in these equations violate the rotational symmetry of Equation 2.2. Enforcement of the rotational symmetry forces one to fit all three components simultaneously, since the dynamics of one component determine the other two uniquely. Up to cubic order there are only two possible terms on the right hand side that transform like vectors.

$$\ddot{\mathbf{x}} = \bar{c}_1 \mathbf{x} + \bar{c}_2 r^2 \mathbf{x} \quad (2.4)$$

This example demonstrates two typical features of enforcing symmetry. (i) The library size has been reduced by more than an order of magnitude. (ii) There are now three times as many measurements that can inform the determination of the coefficients.

A further reason to enforce symmetry is to split up independent information contained in a relation. Consider the Navier-Stokes equation describing the evolution of a flow field  $\mathbf{u}$ .

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \nu \nabla^2 \mathbf{u} \quad (2.5)$$

One can take the gradient of this equation to see the dynamics of the flow gradient matrix  $\mathbf{E} = \nabla \mathbf{u}$ .

$$\partial_t \mathbf{E} + (\mathbf{u} \cdot \nabla) \mathbf{E} + \mathbf{E}^2 + \nabla \nabla p = \nu \nabla^2 \mathbf{E} \quad (2.6)$$

While this system is rotationally symmetric (since it is written in terms of vectors, matrices, and dot products), it contains three independent physical statements. If the flow is incompressible such that  $\text{Tr} \mathbf{E} = \nabla \cdot \mathbf{u} = 0$ , then the trace of the dynamics gives a spatial constraint on the pressure.

$$\text{Tr} (\mathbf{E}^2) + \nabla^2 p = 0 \quad (2.7)$$

If  $\mathbf{E}$  is traceless, it can be written as the sum of a symmetric traceless matrix  $\mathbf{A} = \mathbf{A}^T$  and an antisymmetric matrix  $\mathbf{W} = -\mathbf{W}^T$ . Taking the antisymmetric part of  $\partial_t \mathbf{E}$  gives the

vorticity equation.

$$\partial_t \mathbf{W} + (\mathbf{u} \cdot \nabla) \mathbf{W} + \mathbf{A} \mathbf{W} + \mathbf{W} \mathbf{A} = \nu \nabla^2 \mathbf{W}, \quad (2.8)$$

which is notably independent of the pressure. Lastly, the symmetric trace-free part of  $\partial_t \mathbf{E}$  gives the dynamics of the strain-rate tensor.

$$\partial_t \mathbf{A} + (\mathbf{u} \cdot \nabla) \mathbf{A} + \hat{P}_{TF}(\mathbf{A}^2 + \mathbf{W}^2 + \nabla \nabla p) = \nu \nabla^2 \mathbf{A}, \quad (2.9)$$

where  $\hat{P}_{TF}$  subtracts the trace. This equation is independent of the Laplacian of the pressure. A 9 component PDE has been projected into three independent PDEs with 1, 3, and 5 components, respectively. It is advantageous for machine-learning algorithms to identify these relations independently, regardless of the fact they can be added together to recover Equation 2.6. Learning these equations separately reduces bias and further shrinks the number of unknown coefficients.

## 2.2 Index Notation

Bad notation is often the barrier to mathematical progress. The common vector notations  $\vec{u}$  and  $\mathbf{u}$  are obtuse because they do not naturally distinguish vectors from different vector spaces, and the rank of a tensor is not clear from inspection. A critical step forward for linear algebra is index notation. Each vector space is assigned a unique set of indices, so that the space each vector lives in is obvious by inspection.

$$\begin{aligned} u^i, v^j, w^k &\in U \\ r^A, s^B, t^C &\in V \\ a^\alpha, b^\beta, c^\gamma &\in W \end{aligned} \quad (2.10)$$

The immediate utility of this notation is the presentation of dual vectors. The dual space  $U^*$  of a real vector space  $U$  is the set of linear maps from  $U$  to  $\mathbb{R}$ . Dual vectors can be written with lowered indices:  $u_i \in U^*$ . The action of a dual vector  $u_i \in U^*$  on  $v^i \in U$  is the dot product of their components.

$$u_i[v^i] = \sum_i u_i v^i \equiv u_i v^i \quad (2.11)$$

To simplify tensor equations, one often appeals to the Einstein summation convention: repeated indices imply a sum. Distinct indices imply the tensor product.

$$u_i v^j \equiv \mathbf{u} \otimes \mathbf{v} = \begin{pmatrix} u_1 v^1 & u_1 v^2 & u_1 v^3 & \cdots \\ u_2 v^1 & u_2 v^2 & u_2 v^3 & \cdots \\ u_3 v^1 & u_3 v^2 & u_3 v^3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (2.12)$$

The tensors with arbitrary upper and lower indices can be generated with a sum of tensor products of  $U$  and  $U^*$ . An alternative to index notation is the diagrammatic notation of Penrose and Cvitanović, where upper/lower indices are replaced with directed lines, and different vector spaces have distinct line styles [38, 39, 40]. This is a standard representation for the elementary differentials in Runge-Kutta schemes, where tensors can be linked together via contraction to form trees [41, 42]

The infamous quip “a tensor is anything that transforms like a tensor” arises from the fact that if the elements of  $U$  are acted on by a matrix  $G^i_j$  representing a symmetry operation, then a tensor has the induced transformation

$$T^{ij\cdots}_{kl\cdots} \rightarrow G^i_m G^j_n \cdots T^{mn\cdots}_{op\cdots} G_k^o G_l^p \cdots$$

where  $G_i^j$  is the matrix inverse of  $G^i_j$ :  $G^i_k G_j^k = \delta^i_j$ . Upper indices are often called contravariant and lower indices covariant [43]. The rank of a tensor is  $(p, q)$  where  $p$  is the

number of contravariant indices and  $q$  is the number of covariant indices. There are four important tensor manipulations.

1. Tensor spaces of the same rank form a vector space regardless of index ordering, since different ordered tensor spaces are trivially isomorphic. For example,  $U_{jk}^i \equiv T^i_{jk} + S_j^i{}_k$ .
2. Tensor products produce a tensor of higher rank. For example,  $T^i_{jk} \equiv v^i \omega_j \omega_k$ .
3. A tensor can be contracted, where a sum is performed over a pair of indices from the contravariant and covariant spaces. For example,  $t_i \equiv T^j_{ji}$ . Contraction maps a tensor of rank  $(p, q) \rightarrow (p - 1, q - 1)$ .
4. Indices of a tensor can be permuted. From a tensor  $T_{ij}$  one can construct a second tensor  $T_{ji}$ . Contravariant indices cannot be exchanged with covariant indices as these come from different vector spaces.

Tensors are foundational to modern physics because they ensure covariance under a set of linear transformations (system symmetries). The ‘‘Principle of General Covariance’’ states that if a tensor equation is true in one frame of reference, then it is true in *all frames of reference* [43, 44]. SPIDER searches for physical relations as linear combinations of consistently transforming objects with a specific focus on tensors. Given a set of tensor observables  $\{u^i, \phi, \dots\}$ , tensor libraries of various ranks can be constructed. Feature matrices can then be evaluated by sampling data and relations found with sparse regression. The precise rules for constructing tensor libraries depends on the symmetry group under consideration, and a number of examples will be considered.

As shown in the previous section, it is possible that a tensor equation contains multiple independent physical statements. A guiding principle of SPIDER is that tensor laws should be stated exclusively in irreducible representations of the symmetry group so that each learned equation corresponds to a single irreducible statement. An irreducible representation  $R$  is not isomorphic to the Cartesian product of two independently transforming

sets  $R \neq R_1 \times R_2$ . Representations do not have to have vector space structure, but those considered here will. Examples of irreducible representations are the trace, antisymmetric, and symmetric trace-free tensor spaces we used in the previous section.

The most general projections which will always be relevant are those arising from index permutations. Tensors are free to have their indices permuted, and irreducible representations can be constructed via the Young Tableau [38]. Rank-2 tensors can be split into their symmetric and antisymmetric parts.

$$\begin{array}{ccc}
 T_{ij} & \xrightarrow{\quad\quad\quad} & T_{[ij]} = \frac{1}{2} (T_{ij} - T_{ji}) \\
 & \searrow & \\
 & & T_{(ij)} = \frac{1}{2} (T_{ij} + T_{ji})
 \end{array}$$

This decomposition is unique, and the original tensor can be recovered:  $T_{ij} = T_{[ij]} + T_{(ij)}$ . Further splitting requires the existence of invariant tensors determined by the considered symmetry group. While there is much literature on the existence of irreducible representations of various symmetry groups, we concern ourselves with representations we can construct from available data and differential operators. This constructional paradigm leads us to construct projection operators. Projection operators are symmetry invariant matrices  $\mathbf{P}$  such that  $\mathbf{P}^2 = \mathbf{P}$ .

It is useful to carry out some of this analysis in the diagrammatic notation. Let the identity matrix  $\delta^i_j$  be denoted by a directed line.

$$\delta^i_j \equiv \longrightarrow$$

Symmetrization and antisymmetrization operators  $\delta^{(i}_k \delta^{j)}_l$  and  $\delta^{[i}_k \delta^{j]}_l$  acting on rank 2 tensors can be denoted by white and black boxes, respectively.

$$\begin{array}{c} \boxed{\phantom{00}} \end{array} \longrightarrow = \frac{1}{2} \begin{array}{c} \longrightarrow \\ \longrightarrow \end{array} + \frac{1}{2} \begin{array}{c} \longrightarrow \\ \longrightarrow \end{array}$$

$$\begin{array}{|c} \hline \blacksquare \\ \hline \end{array} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} = \frac{1}{2} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} - \frac{1}{2} \begin{array}{c} \rightarrow \\ \rightarrow \end{array}$$

These are operators acting on rank-2 objects, and it can be checked they are projection operators:  $\mathbf{P}^2 = \mathbf{P}$ .

$$\begin{array}{|c} \hline \square \\ \hline \end{array} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} = \begin{array}{c} \rightarrow \\ \rightarrow \end{array}$$

$$\begin{array}{|c} \hline \blacksquare \\ \hline \end{array} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} = \begin{array}{c} \rightarrow \\ \rightarrow \end{array}$$

Symmetrized and antisymmetrized tensors have no intersection, as the application of both projection operators always results in zero.

$$\begin{array}{|c} \hline \blacksquare \\ \hline \end{array} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} = \begin{array}{c} \rightarrow \\ \rightarrow \end{array} = 0$$

These projection operators have been constructed out of the identity  $\delta^i_j$ , but other symmetry-invariant tensors like the symmetric rank 3 tensor  $d_{ijk}$  of the  $E_6$  family or metric  $g_{ij}$  of Euclidean space should be included when such symmetries are being considered [38]. In fact, any matrix  $\mathbf{M}$  one can construct out of the fundamental tensors of the symmetry group will have invariant subspaces determined by its spectrum. Assuming  $\mathbf{M}$  is diagonalizable with eigenvalues  $\lambda_i$ , projection operators can be constructed explicitly as

$$\mathbf{P}_i = \prod_{j \neq i} \frac{\mathbf{M} - \lambda_j \mathbf{I}}{\lambda_i - \lambda_j} \quad (2.13)$$

where  $\mathbf{I}$  is the identity matrix. These projection operators are unique if all invariant matrices of a system are diagonalizable and commute. The presence of non-commuting invariant tensors destroy uniqueness, but are still useful in projecting relations. An example of this is the index permutation decomposition of rank 3 tensors, where two of the four projection operators are not unique. Symmetrization and antisymmetrization of rank 2 tensors are the



$\nabla_i g_{jk} = \nabla_i g^{jk} = 0$ , which is a standard assumption in Riemannian geometry[45]. Tree tensors need only be computed up to index permutation.  $\mathcal{T}_{(i,j)}$  is the set of tree tensors with  $i$  upper and  $j$  lower indices. The tree tensors up to quadratic order with  $p, q \leq 3$  are

$$\begin{aligned}
\mathcal{T}_{(0,0)} &= \{1\} & \mathcal{T}_{(0,2)} &= \{g_{ij}\} \\
\mathcal{T}_{(1,0)} &= \{v^i\} & \mathcal{T}_{(1,1)} &= \{\nabla_i v^j\} & \mathcal{T}_{(1,2)} &= \{v^i g_{jk}\} \\
\mathcal{T}_{(2,0)} &= \{v^i v^j, g^{ij}\} & \mathcal{T}_{(2,2)} &= \{g_{ij} g^{kl}\} \\
\mathcal{T}_{(3,0)} &= \{v^i g^{jk}\} & & & & 
\end{aligned} \tag{2.14}$$

Although 1 never appears explicitly in the building blocks of tree tensors, it can always appear as a trivial tree tensor as a matter of convention. For some fundamental tensors, there are identities that can be used at the tree tensor stage to simplify subsequent libraries. For example, in  $\text{SO}(3)$  the Levi-Civita tensor satisfies a quadratic identity.

$$\varepsilon_{ijk} \varepsilon^{lmn} \propto \delta^l_i \delta^m_j \delta^n_k \tag{2.15}$$

**2. Compute all possible contractions of tree tensors to produce prelibraries  $\mathcal{P}_{(i,j)}$ .**

Recall that contraction (or trace) is a map  $\text{Tr}$  that reduces the rank of tensors:  $(i, j) \rightarrow (i - 1, j - 1)$ . These sets of contracted objects form pre-libraries. For example, consider the contractions leading to scalars.

$$\mathcal{P}_{(0,0)} = \mathcal{T}_{(0,0)} \cup \text{Tr } \mathcal{T}_{(1,1)} \cup \text{Tr Tr } \mathcal{T}_{(2,2)} = \{1, \nabla_i v^i, g_{ij} g^{ij}\} \tag{2.16}$$

$\mathcal{P}_{(0,0)}$  has an identity since  $g_{ij} g^{ij} = n$ . This is a general feature of library construction. Identities among the building blocks will be embedded in the libraries, and it is possible to recover them and prune them in the same way as nontrivial physical relations. Two options for finding identities are sparse regression on synthetic data or

symbolic manipulation if corresponding algorithms are implemented with sufficient care.

### 3. Project the pre-libraries into irreducible representations to obtain libraries $\mathcal{L}_r$ .

Use the projection operators built from your symmetry group (especially including index permutation) to project the prelibraries into libraries of irreducible representations. There will sometimes be redundancy in representations: for instance  $\mathbf{P}_2$  acting on  $(1, 1)$  tensors will produce scalar equations multiplying  $\delta^i_j$ . No new relations can be found in this  $(1, 1)$  subspace that would not have been found in the scalar library. This  $(1, 1)$  subspace can be safely ignored for the purposes of model discovery.

In the following sections, some examples of symmetry groups will be considered and explicit constructions of low rank irreducible representations will be presented.

### 2.3 The Diffeomorphism Group and $\mathbf{GL}(n, \mathbb{R})$

Many tensors considered in physics are constructed from the tangent space of a manifold. Examples include flow velocities, the electric and magnetic fields, and the gravitational field (both as a metric and Riemannian curvature) [45, 44]. Even the gauge bosons of the electroweak and strong nuclear forces are tensors with one component from the cotangent space [46, 47].

For this reason, a review of the definition of the tangent space will be given in the context of library generation. Let  $M$  be a  $n$ -dimensional manifold and  $p$  be a point on  $M$ . The tangent space  $T_p M$  is the set of directional derivatives of scalar functions at point  $p$ . A vector  $\mathbf{v}$  in the tangent space acts on a scalar function and produces a single real number.

$$\mathbf{v}[\phi] = \sum_{i=1}^n \left( v^i \frac{\partial}{\partial x^i} \phi \right) \Big|_p \quad (2.17)$$

A tangent vector is completely specified by its components  $v^i$ . While  $\mathbf{v}$  is an abstract geometric object with no dependence on coordinates, the values of  $v^i$  will be completely

determined by the choice of coordinates. The cotangent space  $T_p^*M$  is the dual space of  $T_pM$ , and will have lowered indices. The basis of  $T_p^*M$  will be the adjoint basis  $dx^i$ , which is a canonical map of partial derivatives.

$$dx^i \left[ \frac{\partial}{\partial x^j} \right] \equiv \delta_j^i \quad (2.18)$$

A general dual vector  $\omega \in T_p^*M$  will have components defined by  $\omega = \omega_i dx^i$ . The action of a dual vector on a vector becomes a dot product of their components.

$$\omega[\mathbf{v}] = \omega_i dx^i [v^j \partial_j] = \omega_i v^j dx^i [\partial_j] = \omega_i v^j \delta_j^i = \omega_i v^i \quad (2.19)$$

Higher rank tensors can be constructed in the standard way. Tensor fields of a manifold are linear representations of invertible coordinate transformations  $\mathbf{x}'(\mathbf{x})$ , which is effectively the diffeomorphism group – the group of smooth maps from a manifold to itself. The chain rule can be used to obtain the new components of  $\mathbf{v}$ .

$$v^i \frac{\partial}{\partial x^i} = \left( v^i \frac{\partial x'^j}{\partial x^i} \right) \frac{\partial}{\partial x'^j} \quad \rightarrow \quad v'^j = v^i \frac{\partial x'^j}{\partial x^i}, \quad (2.20)$$

with dual vectors transforming via the inverse Jacobian. At each point, the Jacobian will be some element of  $GL(n, \mathbb{R})$ , the group of invertible linear transformation of  $n$  elements. This group has no invariant tensors outside of the identity. Without specific knowledge of the manifold  $M$ , the only guaranteed projections are those accomplished by symmetrization and trace-subtraction.

The partial derivative  $\partial_i$  of a non-scalar is generally not a tensor [43]. If a non-degenerate metric  $g_{ij}$  exists, the Levi-Civita connection defines a covariant derivative<sup>1</sup> that can be used instead [44]. If no natural metric or connection field is present, the only allowed differential operators are the exterior derivative and the Lie derivative [44]. The exterior

---

<sup>1</sup>In Cartesian coordinates with a Euclidean metric,  $\nabla_i = \partial_i$ .

derivative<sup>2</sup> acts on totally antisymmetric covariant tensors  $d\omega = \partial_i\omega_j - \partial_j\omega_i$  and increases the covariant rank by 1. The Lie derivative  $\mathcal{L}_v$  generalizes the directional derivative along a contravariant vector  $v^i$ : it acts on any tensor. For example,  $\mathcal{L}_v\omega_i = v^j\partial_j\omega_i + \omega_j\partial_i v^j$ . These are the only differential operators one is guaranteed on a general manifold that can be used in library construction. The representations up to rank 2 that warrant study are in Table 2.1.

**Table 2.1: Low-dimensional representations of the diffeomorphism group of a general manifold.** Without an invertible quadratic form, there is no natural bijection from covariant to contravariant spaces. For a total rank less than 2, there are 8 distinct irreducible representations physics could live in.

rank	irreducible representations to study	redundant representations
(0,0)	$T$	None
(1,0)	$T^i$	None
(2,0)	$T^{[ij]} \quad T^{(ij)}$	None
(0,1)	$T_i$	None
(0,2)	$T_{[ij]} \quad T_{(ij)}$	None
(1,1)	$T^i_j - \frac{1}{n}\delta^i_j T^k_k$	$\frac{1}{n}\delta^i_j T^k_k$

Suppose there is a manifold with an observable vector field  $v^i$  and an observable one-form  $\omega_i$ . Up to cubic order in  $\{v^i, \omega_i, \partial_i\}$ , there are 10 classes of tree tensors.

$$\begin{aligned}
\mathcal{T}_{(0,0)} &= \{1\} & \mathcal{T}_{(1,0)} &= \{v^i\} & \mathcal{T}_{(2,0)} &= \{v^i v^j\} & \mathcal{T}_{(3,0)} &= \{v^i v^j v^k\} \\
\mathcal{T}_{(0,1)} &= \{\omega_i, \mathcal{L}_v \omega_i\} & \mathcal{T}_{(1,1)} &= \{v^i \omega_j\} & \mathcal{T}_{(2,1)} &= \{v^i v^j \omega_k\} \\
\mathcal{T}_{(0,2)} &= \{\omega_i \omega_j, \partial_{[i} \omega_{j]}\} & \mathcal{T}_{(1,2)} &= \{v^i \omega_j \omega_k, v^i \partial_{[i} \omega_{j]}\} \\
\mathcal{T}_{(0,3)} &= \{\omega_i \omega_j \omega_k, \partial_{[i} \omega_{j]} \omega_k\} & & & & & & 
\end{aligned} \tag{2.21}$$

The lack of structure leads to many sets of tree tensors, but very few elements within them.

<sup>2</sup>The exterior derivative is typically represented in index-free notation as  $d$ . For all totally antisymmetric tensors  $\omega$ ,  $d^2\omega = 0$ .

The only tensor derivatives available are the exterior derivative  $\partial_{[i}\omega_{j]}$  and Lie derivative  $\mathcal{L}_v\omega_i$ <sup>3</sup>. Generating the libraries is done by considering all possible contractions of tree tensors and projecting them to irreducible representations. Contractions map tensors of rank  $(m, n) \rightarrow (m - 1, n - 1)$ , which corresponds to moving diagonally ↖ in the previous collection of equations. The libraries of representations with rank less than 2 is then

$$\begin{aligned}
L_{(0,0)} &= \{1, v^i\omega_i\} \\
L_{(0,1)} &= \{\omega_i, \mathcal{L}_v\omega_i, \omega_i\omega_j v^j, v^j\partial_{[j}\omega_{i]}\} \\
L_{(0,2) \text{ antisymmetric}} &= \{\partial_{[i}\omega_{j]}\} \\
L_{(0,2) \text{ symmetric}} &= \{\omega_i\omega_j\} \\
L_{(1,0)} &= \{v^i, v^i v^j\omega_j\} \\
L_{(1,1) \text{ trace-free}} &= \{v^i\omega_j - \frac{1}{n}\delta_j^i v^k\omega_k\} \\
L_{(2,0) \text{ antisymmetric}} &= \emptyset \\
L_{(2,0) \text{ symmetric}} &= \{v^i v^j\}
\end{aligned} \tag{2.22}$$

These libraries can comprehensively support tensor equations up to cubic in fields and derivatives. This procedure of contraction and projection is completely general, as more specific coordinate symmetry groups can be viewed as introducing further tensor fields (like the metric  $g_{ij}$ ) for the construction of tree tensors.

Note that the diffeomorphism group can generally be any element of  $GL(n, \mathbb{R})$  at a point. If one restricts to a global  $GL(N, \mathbb{R})$  invariance such that there is no nonlinearity in coordinate transformations  $\partial_i\partial_j x^k = 0$ , then the gradient operator  $\partial_i$  becomes a tensor operator. The exterior derivative and Lie derivative are no longer the only acceptable derivatives.

---

<sup>3</sup>Note that  $\mathcal{L}_v v^i = 0$  identically.

## 2.4 O(n)

A theory with O(n) symmetry respects the spatial  $n$ -dimensional Euclidean metric tensor: the symmetry group is the set of rotations and reflections on  $\mathbb{R}^n$ . We will neglect translations and focus on the consequences of invariance with respect to rotations and reflections. The metric tensor  $g_{ij} = g_{ji}$  and its inverse  $g^{ij}$  define the Euclidean dot product, and provide a natural map between vectors and dual vectors. In Cartesian coordinates,  $g_{ij} = g^{ij} = \text{diag}(1, 1, \dots, 1)$  so that two vectors  $\mathbf{u}$  and  $\mathbf{v}$  have a dot product.

$$\mathbf{u} \cdot \mathbf{v} \equiv u^i v^j g_{ij} = u^1 v^1 + u^2 v^2 + \dots \quad (2.23)$$

The metric can be used to raise and lower indices in a consistent way:  $u_i \equiv g_{ij} u^j$  and  $u^i \equiv g^{ij} u_j$ . This puts covariant and contravariant indices on equal footing, and all tensors can be considered fully covariant without loss of generality. Notation can be relaxed to allow for contraction on covariant indices:  $T_{ii} \equiv T_{ij} g^{ij}$ . The effect of this equivalence is that trace-subtraction is now possible for all tensors. The irreducible representations split to no longer have traces anywhere. Table 2.2 shows how this splits rank-2 tensors into a triplet.

$$\begin{array}{ccc} T_{ij} & \xrightarrow{\hspace{10em}} & T_{[ij]} \\ & \searrow & \\ & T_{(ij)} & \xrightarrow{\hspace{10em}} T_{(ij)} - \frac{1}{n} T_{kk} g_{ij} \\ & & \searrow \\ & & T_{kk} g_{ij} \end{array} \quad (2.24)$$

The incompressibility condition and Navier-Stokes equations are an example of O(n) invariance describing a constrained flow field  $u_i$  and pressure  $p$ .

$$\nabla_i u_i = 0 \quad (2.25)$$

$$\partial_t u_i + u_j \nabla_j u_i + \nabla_i p - \nu \nabla^2 u_i = 0 \quad (2.26)$$

Table 2.2: **The irreducible tensors up to rank 2 of the  $O(n)$  symmetry group.** The rank-2 tensors split into three irreducible tensor spaces, however the representation proportional to the metric is redundantly isomorphic to the scalar library.

rank	irreducible representations to study	redundant representations
0	$T$	None
1	$T_i$	None
2	$\mathbf{P}^{AS}T_{ij} = \frac{1}{2}(T_{ij} - T_{ji})$ $\mathbf{P}^{STF}T_{ij} = \frac{1}{2}(T_{ij} + T_{ji} - \frac{2}{n}g_{ij}T_{kk})$	$\frac{2}{n}g_{ij}T_{kk}$

where  $\nu$  is the kinematic viscosity and  $\nabla^2 \equiv \nabla_i \nabla_i$ . The metric  $g_{ij}$  implicitly enters into these equations in three places: to define a covariant derivative and therefore the divergence of  $u_i$ , to allow the dual vector  $\nabla_i p$  to couple to the naturally contravariant  $u^i$ , and for the existence of dissipation by  $\nabla^2$ . It is instructive to construct libraries for this system up to cubic in  $\{u_i, p, \nabla_i, \partial_t\}$ . There is never a reason to multiply by  $g_{ij}$  in tree tensor construction, since the only additional tensors this will produce will be projected into redundant representations.

$$\begin{aligned}
\mathcal{T}_0 &= \{1, p, \partial_t p, \partial_t^2 p, p^2, p^3, p \partial_t p\} \\
\mathcal{T}_1 &= \{u_i, p u_i, p^2 u_i, \partial_t p u_i, \partial_t u_i, \partial_t^2 u_i, \nabla_i p, \partial_t \nabla_i p\} \\
\mathcal{T}_2 &= \{u_i u_j, p u_i u_j, p \nabla_i u_j, \nabla_i u_j, \nabla_i p u_j, \nabla_i \nabla_j p, \partial_t u_i u_j\} \\
\mathcal{T}_3 &= \{u_i u_j u_k, u_i \nabla_j u_k, \nabla_i \nabla_j u_k\}
\end{aligned} \tag{2.27}$$

The prelibraries are computed by taking all possible contractions.  $\mathcal{P}_3$  is ignored since we only consider tensors of rank 2 or less for model discovery.

$$\begin{aligned}
\mathcal{P}_0 &= \{p, \partial_t p, \partial_t^2 p, p^2, p^3, p \partial_t p, u^2, p u^2, \nabla_i u_i, p \nabla_i u_i, u_i \nabla_i p, \nabla^2 p, u_i \partial_t u_i\} \\
\mathcal{P}_1 &= \{u_i, p u_i, p^2 u_i, \partial_t p u_i, \partial_t u_i, \partial_t^2 u_i, \nabla_i p, \partial_t \nabla_i p, u^2 u_i, u_i (\nabla_j u_j), u_j \nabla_j u_i, \nabla_i \nabla_j u_j, \nabla^2 u_i\} \\
\mathcal{P}_2 &= \{u_i u_j, p u_i u_j, \nabla_i u_j, p \nabla_i u_j, \nabla_i p u_j, \nabla_i \nabla_j p, \partial_t u_i u_j\}
\end{aligned} \tag{2.28}$$

The only projections to do for  $O(n)$  are the symmetric traceless and antisymmetric projec-

tions on rank 2 tensors.

$$\begin{aligned}
\mathcal{L}_0 &= \{1, p, \partial_t p, \partial_t^2 p, p^2, p^3, p\partial_t p, u^2, pu^2, p\nabla_i u_i, u_i \nabla_i p, \nabla^2 p, u_i \partial_t u_i\} \\
\mathcal{L}_1 &= \{u_i, pu_i, p^2 u_i, \partial_t pu_i, \partial_t u_i, \partial_t^2 u_i, \nabla_i p, \partial_t \nabla_i p, u^2 u_i, u_i (\nabla_j u_j), u_j \nabla_j u_i\} \\
\mathcal{L}_2^{AS} &= \mathbf{P}^{AS} \{\nabla_i u_j, p\nabla_i u_j, \nabla_i pu_j, \partial_t u_i u_j\} \\
\mathcal{L}_2^{STF} &= \mathbf{P}^{STF} \{u_i u_j, pu_i u_j, \nabla_i u_j, p\nabla_i u_j, \nabla_i pu_j, \nabla_i \nabla_j p, \partial_t u_i u_j\} \tag{2.29}
\end{aligned}$$

$\mathcal{L}_0$  in its current form is not rich enough to discover the energy and pressure Poisson equations, which are obtained by taking the contraction of Equation 2.26 with  $u_i$  and  $\nabla_i$ , respectively:

$$\nabla^2 p + \nabla_i (u_j \nabla_j u_i) = 0, \tag{2.30}$$

$$\partial_t E + u_i \nabla_i E + u_i \nabla_i p + \nu \nabla^2 E - (\nu \nabla_i u_j) (\nabla_i u_j) = 0, \quad E = \frac{1}{2} u^2. \tag{2.31}$$

To this end, we introduce quartic terms from the additional rank 4 tree tensors  $\nabla_i u_j \nabla_k u_l$  and  $u_i \nabla_j \nabla_k u_l$  and a rank 0 term  $u_i u_j \nabla_i u_j$  corresponding to energy advection. Since these are rank 4 tree tensors, scalars can be obtained from double contraction.

$$\mathcal{P}_0 \rightarrow \mathcal{P}_0 \cup \{(\nabla_i u_i)^2, \nabla_i u_j \nabla_i u_j, \nabla_i u_j \nabla_j u_i, u_i \nabla^2 u_i, u_i \nabla_i \nabla_j u_j, u_i u_j \nabla_i u_j\} \tag{2.32}$$

$$\begin{aligned}
\mathcal{P}_2 \rightarrow \mathcal{P}_2 \cup \{ &\nabla_i u_j \nabla_k u_k, \nabla_i u_k \nabla_k u_j, \nabla_k u_i \nabla_k u_j, \nabla_j u_k \nabla_i u_k, \\ &u_i \nabla_j \nabla_k u_k, u_k \nabla_i \nabla_j u_k, u_k \nabla_k \nabla_i u_k\} \tag{2.33}
\end{aligned}$$

These new elements of  $\mathcal{P}_2$  will need to be projected into the antisymmetric and symmetric trace-free spaces.

Note that the cross product induced by  $\varepsilon_{ijk}$  is not left invariant by the  $O(n)$  group. No cross-products are needed in any symmetries. It is instructive to show Maxwell's equations need no cross product to be written covariantly. One takes the magnetic field  $B_{ij} \equiv \varepsilon_{ijk} B_k$

to be an antisymmetric tensor instead of a vector. Then Maxwell's equations become

$$\nabla_i E_i = \rho, \quad \nabla_i B_{jk} + \nabla_j B_{ki} + \nabla_k B_{ij} = 0, \quad (2.34)$$

$$\partial_t E_i = \nabla_j B_{ij} - j_i, \quad \partial_t B_{ij} = \nabla_j E_i - \nabla_i E_j. \quad (2.35)$$

In this sense, the relativistic formulation of electromagnetism in terms of a rank-2 tensor  $F_{\mu\nu}$  is not a far leap from a  $O(3)$  covariant formulation.

As a final comment on  $O(n)$  symmetry, library construction proceeds exactly the same for pseudo-orthogonal groups  $O(m,n)$ , which preserve the quadratic form  $x_1^2 + \dots + x_m^2 - y_1^2 - \dots - y_n^2$ . As long as the metric tensor is invertible, all projection operators can be constructed identically. This is of particular interest to relativistic physics with Lorenz symmetry. There the Minkowski metric  $\eta^{\alpha\beta}$  is the invariant of  $O(3,1)$ .

## 2.5 $SO(n)$

$SO(n)$  is a subgroup of  $O(n)$  that excludes reflections. Tensors can be constructed out of the totally antisymmetric rank- $n$  orientation tensor  $\varepsilon_{ij\dots k}$ . The metric  $g_{ij}$  can again be used to keep all indices covariant. Since the rank of this orientation tensor depends on  $n$ , the constructable projection operators will be quite different for each  $n$ . Physically, theories with only  $SO(n)$  invariance are not agnostic to reflections: the physics may depend on the handedness of structures.

The major tool that the orientation form introduces is the Hodge dual or Hodge star operator  $\star$  [44]. This provides a bijection between antisymmetric tensors of rank  $r$  and  $n - r$  by contraction of all indices with  $\varepsilon_{ij\dots k}$ .  $SO(2)$  has the orientation tensor  $\varepsilon_{ij}$  with two important consequences for tensor library construction. The first is that any antisymmetry can be factored out of a tensor. This follows from the quadratic identity

$$\varepsilon_{ij}\varepsilon_{kl} = \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk} \quad (2.36)$$

Applying this to a tensor  $F_{ij\dots} = -F_{ji\dots}$  one sees that any tensor with antisymmetry in two indices is proportional to  $\varepsilon_{ij}$ .

$$F_{ij\dots} = \frac{1}{2}\varepsilon_{ij}\varepsilon_{kl}F_{kl\dots} \quad (2.37)$$

This statement is independent of the metric and only depends on the orientation form (so it holds in  $\text{SL}(2, \mathbb{C})$  as well) [48]. There is no reason to consider tensors with *any antisymmetry* under the action of  $\text{SO}(2)$ . Any physical relations with antisymmetry can always be captured instead with a lower-rank representation of the group. The second important aspect of  $\text{SO}(2)$  is that the Hodge dual of a vector is a vector: the Hodge dual is a matrix operator usable in projections

$$\star \star v_i = \varepsilon_{ij}\varepsilon_{jk}v_k = -v_i \quad (2.38)$$

The eigenvalues of the Hodge star are  $\pm i$ , so projection can only be accomplished by allowing coefficients to be complex. The projection operators are straightforward to construct.

$$v_i^\pm = v_i \pm i \star v_i \quad \star v_i^\pm = \pm i v_i^\pm \quad (2.39)$$

The rank 1 space is split into one-dimensional complex representations of  $\text{SO}(2)$ , known as self-dual  $v_i^+$  and anti-self-dual  $v_i^-$  representations. These spaces are more simply described as the action of  $\text{U}(1)$  on  $\mathbb{C}$ , where  $v^\pm = v_x \pm i v_y$ . Tensors of rank  $n$  transform like

$$T_n \rightarrow \exp(in\phi)T_n$$

when  $v^+ \rightarrow \exp(i\phi)v^+$ . One can see that complex conjugate maps tensors of rank  $n$  to rank  $-n$ , and complex scalar multiplication is equivalent to tensor multiplication. The spatial derivative can be promoted to a  $\text{U}(1)$  operator of rank 1 as well  $D \equiv \partial_x + i\partial_y$ .

Library generation is amazingly simple for  $\text{U}(1)$  symmetric theories. When constructed in the complex form, there is no need for projections or contractions. The tree tensors

themselves are immediately the libraries. Also, no negative rank libraries need construction, since these relations will be complex conjugates of positive rank relations. Table 2.6 presents the irreducible representations to study for  $\text{SO}(2) \cong \text{U}(1)$ .

Table 2.3: **The irreducible tensors up to rank 2 of the  $\text{SO}(2) \cong \text{U}(1)$  symmetry group.** It is an interesting feature of this group that only one non-redundant representation exists for each rank. Tensor multiplication for these 1D  $\mathbb{C}$  spaces is just ordinary complex multiplication.

abs(rank)	irreducible representations to study	redundant representations
0	$T$	$T^+ \otimes T^-$
1	$T^+$	$T^-$
2	$T^+ \otimes T^+$	$T^- \otimes T^-$

A well-known example of a theory with  $\text{U}(1)$  invariance is the Schrödinger equation with the invariant  $\psi\bar{\psi}$  being associated with a probability density. Although this symmetry is usually decoupled from spatial rotations, consider the wavefunction  $\psi(x, y, t)$  to be a rank 1 self-dual vector under the action of  $\text{U}(1)$  induced by  $\text{SO}(2)$ . The differential operators of the theory are  $\{\partial_t, D, \bar{D}\}$  or ranks 0, 1, and -1, respectively. The only field we will consider is  $\psi$ , which is rank 1. The tree tensors up to rank 2 are themselves the libraries worth studying. The libraries are presented up to cubic order in operators and fields.

$$\mathcal{T}_0 = \mathcal{L}_0 = \{1, \psi\bar{\psi}, \psi\partial_t\bar{\psi}, \partial_t\psi\bar{\psi}, D\bar{\psi}, \bar{D}\psi, \partial_t D\bar{\psi}, \partial_t \bar{D}\psi\} \quad (2.40)$$

$$\mathcal{T}_1 = \mathcal{L}_1 = \{\psi, \partial_t\psi, \partial_t^2\psi, D\bar{D}\psi, \psi^2\bar{\psi}, D^2\bar{\psi}, \psi D\bar{\psi}, \bar{\psi} D\psi\} \quad (2.41)$$

$$\mathcal{T}_2 = \mathcal{L}_2 = \{\psi^2, \psi\partial_t\psi, D\psi, \partial_t D\psi\} \quad (2.42)$$

For instance, the free-particle Schrödinger equation contains terms from the library  $\mathcal{L}_1$  with complex coefficients.

$\text{SO}(3)$  has an orientation form  $\varepsilon_{ijk}$ . For odd dimensions, the Hodge dual does not map any tensor to a same rank tensor, so there is no notion of self-dual within a single tensor rank. This can be overcome by considering direct sums of different tensor spaces.

For example,  $T_i \oplus T_{jk}$  with  $T_{jk} = -T_{kj}$  is closed under the Hodge dual: it interchanges these two spaces. The dual operation can be diagonalized on this space, but there is no demonstrable economy offered by such a formalism for model discovery.

The true benefit of SO(3) symmetry is that one no longer needs to study the antisymmetric rank-2 space. It is redundant since its Hodge dual lives in the rank-1 space. Other than including  $\varepsilon_{ijk}$  in tree tensors, there is no subtlety in library construction for SO(3) theories. For example, the vector library of the fluid system discussed with O(n) symmetry will be  $\mathcal{L}_1 \cup \varepsilon[\mathcal{L}_{AS}]$ :

$$\begin{aligned} \mathcal{L}_1 = & \{u_i, pu_i, p^2u_i, \partial_t pu_i, \partial_t u_i, \partial_t^2 u_i, \nabla_i p, \partial_t \nabla_i p, u^2 u_i, u_i(\nabla_j u_j), u_j \nabla_j u_i\} \\ & \cup \{\varepsilon_{ijk} \nabla_j u_k, \varepsilon_{ijk} p \nabla_j u_k, \varepsilon_{ijk} \nabla_j p u_k, \varepsilon_{ijk} \partial_t u_j u_k\}. \end{aligned} \quad (2.43)$$

This addition of antisymmetric tensors contracted with  $\varepsilon_{ijk}$  is equivalent to the introduction of pseudovectors. Pseudovectors transform like vectors under rotations but change sign under reflections. For O(n) symmetry, no such coupling is possible. Contracting  $\varepsilon_{ijk}$  with a totally antisymmetric tensor  $T_{ijk}$  produces pseudoscalars that can now enter the scalar library. As a further example, Maxwell's equations can be written in vector form only when SO(3) symmetry is assumed. Then the magnetic field  $B_i \equiv 1/2 \varepsilon_{ijk} B_{jk}$  can be used to write the equations as

$$\nabla_i E_i = \rho, \quad \nabla_i B_i = 0, \quad (2.44)$$

$$\partial_t E_i = \varepsilon_{ijk} \nabla_j B_k - j_i, \quad \partial_t B_i = -\varepsilon_{ijk} \nabla_j E_k. \quad (2.45)$$

Table 2.4: **The irreducible tensors up to rank 2 of the SO(3) symmetry group.** Like SO(2), there is a single representation per rank warranting investigation.

rank	irreducible representations to study	redundant representations
0	$T$	None
1	$T_i$	None
2	$\frac{1}{2} (T_{ij} + T_{ji} - \frac{2}{n} g_{ij} T_{kk})$	$\frac{2}{n} g_{ij} T_{kk}, \frac{1}{2} (T_{ij} - T_{ji})$

An important final remark is that for tree tensor construction of  $\text{SO}(n)$  theories,  $\varepsilon_{ij\dots l}$  only needs to be included up to once per tree tensor. This is because regardless of dimension, the tensor square of the orientation form is proportional to a tensor product of identities [38].

## 2.6 The Homogeneous Galilean Group

Galilean invariance is the symmetry group of particles moving at nonrelativistic speeds interacting via Newton's laws. It is in particular the symmetry group of particles with interactions given by the standard Hamiltonian

$$H(\mathbf{q}_a, \mathbf{p}_a) = \sum_a \frac{\mathbf{p}_a \cdot \mathbf{p}_a}{2m_a} + \sum_{a \neq b} V(\|\mathbf{q}_a - \mathbf{q}_b\|_2), \quad (2.46)$$

where  $a$  is an index over particles. We will restrict our attention to the homogeneous Galilean group, in which there is no translation in space or time [49]. An element  $g$  of the homogeneous Galilean group  $G$  corresponds to a rotation  $\hat{R}$  and a boost  $\vec{v}$  of the spatial components.

$$g(\vec{x}, t) = (\hat{R}\vec{x} - \vec{v}t, t) \quad (2.47)$$

Representations of the Galilean group have been the subject of extensive investigation: unitary representations for quantum systems have been developed [50], as well as four-dimensional representation with a degenerate metric [51, 52] and a five-dimensional representations with a non-degenerate metric [53]. The latter two theories are of interest to this work as they give guidelines for constructing indexed representations from Galilean-covariant data. Galilean covariant schemes for machine learning have been developed [11, 10], but in that work only a particular irreducible representation was considered.

The four-dimensional representation concerns itself with four-vectors  $x^\mu = (x, y, z, t)$  just like special relativity. The Galilean group is the set of transformations preserving a

dual vector  $\omega_\alpha$  and a degenerate metric  $h^{\alpha\beta}$ .

$$\omega_\alpha = (0, 0, 0, 1) \quad h^{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (2.48)$$

No inverse of  $h^{\alpha\beta}$  exists, so the vector  $v^\alpha$  has no Galilean invariant norm. The inner product is only defined for covariant indices. Indices of dual vectors in  $V^*$  can be raised by  $h^{\alpha\beta}$ , but contravariant vectors in  $V$  have no map back to  $V^*$ . The invariant dual vector  $\omega_\alpha$  makes the Galilean group unlike any other group discussed in this work. Note that  $h^{\alpha\beta}\omega_\beta = 0$ , so that no projection matrix can be constructed.  $\omega_\alpha$  defines an invariant subspace  $U = \{x^\alpha \mid x^\alpha \in V, x^\alpha\omega_\alpha = 0\}$ . The subspace  $U$  consists of the vectors with vanishing time component, and the action of the Galilean group on this subspace is simply the orthogonal group. This space has its own dual space  $U^*$  that **is not isomorphic** to the dual space  $V^*$ . Let vectors in  $U$  have indices labeled with Latin letters  $a, b, \dots$ . Then  $U$  has its own invariant tensors: the Euclidean metric  $\gamma_{ab}$  and its inverse  $\gamma^{ab}$  providing a bijection between  $U$  and  $U^*$ . The inclusion map  $\iota : U \rightarrow V$  can be defined by changing an upper Latin index to a Greek index:  $\iota(x^a) = \iota_b^\alpha x^b = x^\alpha$ . The relations between these spaces are summarized in Equation 2.49.

$$\begin{array}{ccc} U & \xrightarrow{\iota_b^\alpha} & V \\ \begin{array}{c} \uparrow \gamma^{ab} \\ \downarrow \gamma_{ab} \end{array} & & \uparrow h^{\alpha\beta} \\ U^* & & V^* \end{array} \quad (2.49)$$

General tensors can then have four kinds of indices:  $T^\alpha_\beta{}^c{}_d$ . All indices can be mapped to  $V$  and  $V^*$  such that all tensors have a canonical form  $T^{\alpha\dots}_\beta\dots$ , so that library construction only needs to concern itself with the covariant and contravariant indices. The unresolved

problem with library construction is that projection operators cannot be constructed naturally. The Galilean invariant matrix  $\mathbf{M} = g^{\alpha\beta}\omega_\gamma\omega_\epsilon$  acting on rank 2 tensors has the unfortunate characteristic equation  $\mathbf{M}^2 = 0$ .  $\mathbf{M}$  is not diagonalizable so Equation 2.13 fails to produce projection operators. It is possible to get around this by introducing an observer field  $v^\alpha$  with  $v^\alpha\omega_\alpha = 1$  such that Galilean transformations induce Milne boosts [52]:

$$v^\alpha \rightarrow v^\alpha + h^{\alpha\beta}u_\beta, \quad (2.50)$$

where  $u_\beta$  is arbitrary. It is clear that  ${}^{(1)}\mathbf{P}_\beta^\alpha = h^{\alpha\beta}\omega_\beta$  and  ${}^{(2)}\mathbf{P}_\beta^\alpha = \delta_\beta^\alpha - h^{\alpha\beta}\omega_\beta$  are projection operators, which map tensors into subspaces that transform with consistent Milne boosts.  ${}^{(2)}\mathbf{P}_\beta^\alpha$  is a projection from  $V$  to  $U$ , such that the Euclidean metric can be used. This is a circumstance where projection may need to be done additionally before contraction of tree tensors so the Euclidean metric can enter. In a study of Reynolds averaged turbulence, Galilean invariance was enforced by working with tensors a priori in the spatial subspace  $U$  to allow dot products [10]. More work is needed to define the rules for library generation, perhaps by moving to the 5D formalism with an invertible metric where the Galilean group is a subgroup of the pseudo-orthogonal group  $O(4,1)$  [53]. Regardless, the representations of  $GL(n, \mathbb{R})$  given in Table 2.1 are good enough to discover any Galilean invariant physics as long as  $\omega_\alpha$  and  $h^{\alpha\beta}$  are included as observables during library construction.

## 2.7 $SU(2)$ and $SL(2, \mathbb{R})$ spinors

It is not the case that all physically relevant tensors are constructed from the tangent space and its dual. One of the most important exceptions in physics is the spinor: a 2-component complex vector  $\psi^A \in \mathbb{C}^2$  with  $A = 0, 1$ . In fact, any theory invariant under the action of  $SO(3)$  has a natural embedding in the language of  $SU(2)$  spinors, which can be helpful for finding solutions in  $SU(2)$  Yang-Mills theory [54, 55]. While spinors have been used extensively in modeling quantum systems, they have applications to classical relativistic

fields [48].

Spinors are complex vectors, so there is a need to define its complex conjugate  $\bar{\psi}^{A'}$ , where  $A'$  is a conjugate index. For a linear transformation  $\Lambda^A_B$  on spin space, there are induced linear transformations on the conjugate and dual spaces.

$$\begin{aligned}\psi^A &\rightarrow \Lambda^A_B \psi^B & \phi_A &\rightarrow (\Lambda^{-1})^B_A \phi_B \\ \bar{\psi}^{A'} &\rightarrow \bar{\Lambda}^{A'}_{B'} \bar{\psi}^{B'} & \bar{\phi}_{A'} &\rightarrow (\bar{\Lambda}^{-1})^{B'}_{A'} \bar{\phi}_{B'}\end{aligned}\quad (2.51)$$

A generic spin tensor can therefore have four kinds of indices:  $T^A_B{}^{A'}_{B'}$ . The indices  $A$  and  $A'$  are unrelated, although one can usually assume they exchange under complex conjugation.  $SU(2)$  is defined by its invariants: the Hermitian norm  $\delta_{AA'} = \text{diag}(1, 1)$  and the orientation form  $\varepsilon_{AB} = -\varepsilon_{BA}$  with  $\varepsilon_{01} = 1$ . These both have inverses  $\delta_{AA'} \delta^{BA'} = \delta^B_A$  and  $\varepsilon_{AC} \varepsilon^{BC} = \delta^B_A$ . These tensors define a natural linear bijection from an arbitrary spinor to a spinor with a lower unconjugated index.

$$\begin{aligned}\psi^A &\rightarrow \psi_A \equiv \varepsilon_{AB} \psi^B \\ \phi^{A'} &\rightarrow \phi_A \equiv \delta_{AA'} \phi^{A'} \\ \chi_{A'} &\rightarrow \chi_A \equiv \varepsilon_{AB} \delta^{BB'} \chi_{B'}\end{aligned}\quad (2.52)$$

These operations can be applied in parallel to an arbitrary number of indices, so that any spin tensor can be made to have exclusively lower unconjugated indices. Only these objects need to be studied to understand  $SU(2)$  covariant physics. Note that repeated conjugation leads to a change in sign:  $\bar{\bar{\psi}}_A = -\psi_A$ . For a rank  $n$  spin tensor, repeated conjugation will change the sign by  $(-1)^n$ . The Hermitian norm  $\delta_{AA'}$  never appears in this notation explicitly and can be discarded from our lexicon. Given a tensor with multiple indices like  $T_{AB}$ , a contraction can be defined by an implicit contraction with  $\varepsilon^{AB}$ .

$$T_{AA} \equiv \varepsilon^{AB} T_{AB} = T_{01} - T_{10}\quad (2.53)$$

Care is required when contracting indices from distinct tensors as the order of multiplication now matters:  $S_{A\dots}T_{A\dots} = -T_{A\dots}S_{A\dots}$ . This is an irrelevant detail from the perspective of library generation since the order of tensors will only change the signs of coefficients in any governing equations.

When the action of  $SU(2)$  is tied to rotations of physical space, as it is in the theory of the electron, it is useful to establish the natural embedding of Euclidean  $\mathbb{R}^3$  in spin space. Rank-2 tensors  $T_{AB}$  can immediately be split into symmetric and antisymmetric parts:  $T_{AB} = T_{(AB)} + T_{[AB]}$ . The antisymmetric tensors are simply a scalar  $T_{AA}$  multiplying  $\varepsilon_{AB}$  and require no further investigation [48].

$$T_{[AB]} = \frac{1}{2}\varepsilon_{AB}T_{CC} \quad (2.54)$$

The symmetric subspace  $T_{(AB)}$  has no trace by symmetry  $T_{(AA)} = 0$ , so there is no linear projection into further complex invariant subspaces. There is however a real subspace with respect to complex conjugation.

$$T_{AB} = \begin{pmatrix} T_{00} & T_{01} \\ T_{01} & T_{11} \end{pmatrix} \quad \bar{T}_{AB} = \begin{pmatrix} \bar{T}_{11} & -\bar{T}_{01} \\ -\bar{T}_{01} & \bar{T}_{00} \end{pmatrix} \quad (2.55)$$

Following Penrose, we define a world vector to be a symmetric rank-2 tensor  $S_{AB} = S_{BA}$  that is equal to its conjugate  $S_{AB} = \bar{S}_{AB}$ . The identification with Euclidean space can be made using  $(x, y, z) \in \mathbb{R}^3$ .

$$S_{AB} = \begin{pmatrix} x + iy & iz \\ iz & x - iy \end{pmatrix} \quad S_{AB}S_{AB} = 2(x^2 + y^2 + z^2) \quad (2.56)$$

3D Euclidean space is therefore an invariant subspace of rank-2 spin tensors. Every vector field  $\vec{v}$  is equivalent to a world vector  $v_{AB}$ . In fact, one can extend the gradient operator  $\vec{\nabla}$

to spin space by defining  $\nabla_{AB}$  with the world vector properties.

$$\nabla_{AB} \equiv \begin{pmatrix} \partial_x + i\partial_y & i\partial_z \\ i\partial_z & \partial_x - i\partial_y \end{pmatrix} \quad (2.57)$$

The gradient being a rank-2 operator fundamentally changes the kinds of equations that can exist. It is possible to take a half-divergence, such that only one of the gradient indices is contracted. This half-divergence naturally defines the 3D divergence and curl of a vector field simultaneously.

$$\nabla_{AC}v_{CB} = -(\vec{\nabla} \times \vec{v})_{AB} + (\nabla \cdot \vec{v})\varepsilon_{AB} \quad (2.58)$$

The usual 3D vector calculus is therefore naturally embedded in SU(2) spinor calculus. To see library construction, consider a Weyl spinor  $\psi_A$ .

$$\partial_t\psi_A + \nabla_{AB}\psi_B = 0 \quad (2.59)$$

The fundamental tensors of this system are  $\{\psi_A, \bar{\psi}_A\}$  and the differential operators are  $\{\nabla_{AB}, \partial_t\}$  of rank 2 and 0. If we allow all tree tensors up to cubic order, the maximal rank will be rank 5.

$$\begin{aligned} \mathcal{T}_0 &= \{1\} \\ \mathcal{T}_1 &= \{\psi_A, \partial_t\psi_A, \partial_t^2\psi^A, \bar{\psi}_A, \partial_t\bar{\psi}_A, \partial_t^2\bar{\psi}^A\} \\ \mathcal{T}_2 &= \{\psi_A\psi_B, \bar{\psi}_A\bar{\psi}_B, \psi_A\partial_t\psi_B, \bar{\psi}_A\partial_t\bar{\psi}_B, \psi_A\partial_t\bar{\psi}_B, \bar{\psi}_A\partial_t\psi_B\} \\ \mathcal{T}_3 &= \{\psi_A\psi_B\psi_C, \bar{\psi}_A\bar{\psi}_B\bar{\psi}_C, \bar{\psi}_A\bar{\psi}_B\psi_C, \bar{\psi}_A\bar{\psi}_B\bar{\psi}_C, \nabla_{AB}\psi_C, \nabla_{AB}\bar{\psi}_C, \partial_t\nabla_{AB}\psi_C, \partial_t\nabla_{AB}\bar{\psi}_C\} \\ \mathcal{T}_4 &= \{\psi_A\nabla_{BC}\psi_D, \bar{\psi}_A\nabla_{BC}\bar{\psi}_D, \psi_A\nabla_{BC}\bar{\psi}_D, \bar{\psi}_A\nabla_{BC}\psi_D\} \\ \mathcal{T}_5 &= \{\nabla_{AB}\nabla_{CD}\psi_E, \nabla_{AB}\nabla_{CD}\bar{\psi}_E\} \end{aligned} \quad (2.60)$$

$\varepsilon_{AB}$  is not needed in constructing tree tensors; if it is not killed by contraction it will lead to a tensor in a redundant representation. Contraction reduces the rank of tensors by 2, so  $\mathcal{L}_0$  can be constructed from traces of  $\mathcal{T}_2$  and  $\mathcal{T}_4$ .

$$\begin{aligned} \mathcal{L}_0 = \{ & \bar{\psi}_A \psi_A, \psi_A \partial_t \psi_A, \bar{\psi}_A \partial_t \psi_A, \psi_A \partial_t \bar{\psi}_B, \bar{\psi}_A \partial_t \bar{\psi}_A \\ & \psi_A \nabla_{AB} \psi_B, \bar{\psi}_A \nabla_{AB} \psi_B, \psi_A \nabla_{AB} \bar{\psi}_B, \bar{\psi}_A \nabla_{AB} \bar{\psi}_B \} \end{aligned} \quad (2.61)$$

Table 2.5: **The irreducible tensors up to rank 2 of the SU(2) symmetry group.**

rank	irreducible representations to study	redundant representations
0	$T$	None
1	$T_A$	None
2	$T_{(AB)}$	$T_{[AB]}$

Just as SU(2) spinors have a natural group action under continuous rotations of SO(3), SL(2,ℂ) spinors have a natural group action under continuous Lorentz transformations of SO(3,1). SL(2,ℂ) only leaves  $\varepsilon_{AB}$  and its inverse/complex conjugate invariant. This means there is no map from conjugate indices to unconjugated ones, so a general spin tensor can be projected to have all covariant indices  $T_{AB\dots A'B'\dots}$ . One can take the complex conjugate of an equation so that the number of unconjugated indices is greater than or equal to the number of conjugated indices. The representations worth studying are identical to the SU(2) case, with the addition of a mixed tensor space  $T_{AA'}$ . This new space contains a real 4D subspace  $\bar{T}_{A'A} = T_{AA'}$  that can be identified with Minkowski spacetime [48].

$$x_{AA'} = \begin{pmatrix} t+z & x+iy \\ x-iy & t-z \end{pmatrix} \quad x_{AA'} x_{BB'} \varepsilon^{AB} \varepsilon^{A'B'} = 2(t^2 - x^2 - y^2 - z^2) \quad (2.62)$$

This natural emergence of the Minkowski metric is why particle physicists prefer a time-positive metric signature. The derivative likewise becomes a rank (1,1) object  $\nabla_{AA'}$ .

**Table 2.6: The irreducible tensors up to rank 2 of the  $SL(2, \mathbb{C})$  symmetry group describing relativistic spinors.**

rank	irreducible representations to study	redundant representations
(0,0)	$T$	None
(1,0)	$T_A$	None
(0,1)	None	$T_{A'}$
(1,1)	$T_{AA'}$	None
(2,0)	$T_{(AB)}$	$T_{[AB]}$

## CHAPTER 3

### FEATURE MATRIX CONSTRUCTION AND SPARSE REGRESSION

Model discovery via regression requires a feature matrix<sup>1</sup>  $G$ , which is constructed by evaluating a library of terms  $\mathcal{L} = \{x, x^2, \sin(x), \dots\}$  at various spatiotemporal locations. Regression then fits a vector  $\mathbf{g}$  corresponding to the dynamics  $\dot{x}$  to obtain a sparse coefficient vector  $\mathbf{g} \approx G\mathbf{c}$ . Estimating the derivative  $\dot{x}$  at a point with finite differences is sensitive to noise, which is certainly present in any real world data. In Rudy et al.’s description of the PDE-find algorithm [56], noise levels above 1% could ruin model discovery for common PDEs including Navier-Stokes in vorticity form. It has been shown that the weak formulation is significantly more noise robust [35, 34, 25, 26] and suitable for handling real experimental data [36, 37].

#### 3.1 Matrix Construction

SPIDER always uses the weak formulation of differential equations, in which elements of  $G$  correspond to numerical integrals of library terms over spacetime domains  $\Omega_i$  rather than evaluations at single points. Considering now fields with spatiotemporal<sup>2</sup> variation  $\mathbf{u}(\mathbf{x})$ , the library can contain nonlinear functions  $\mathcal{L} = \{\mathbf{f}_j(\mathbf{x}, \mathbf{u}, \nabla\mathbf{u}, \dots)\}$ . The feature matrix  $G$  is formally defined by

$$G_{ij} = \frac{1}{N_i S_j} \int_{\Omega_i} dV w_i f_j \quad (3.1)$$

$$N_i = \int_{\Omega_i} dV |w_i| \quad (3.2)$$

---

<sup>1</sup>This matrix has gone by many names in the literature. SINDy papers tend to use  $\Theta$  and precursors of the SPIDER papers use  $Q$ . A large amount of my time at Georgia Tech has been devoted to analyzing an active nematic system with an observable field  $Q_{ij}$ . To avoid confusion in that case, I have introduced another convention  $G$ .

<sup>2</sup> $\mathbf{x}$  is being used as a spacetime vector  $\mathbf{x} = \langle t, x, y, z, \dots \rangle$

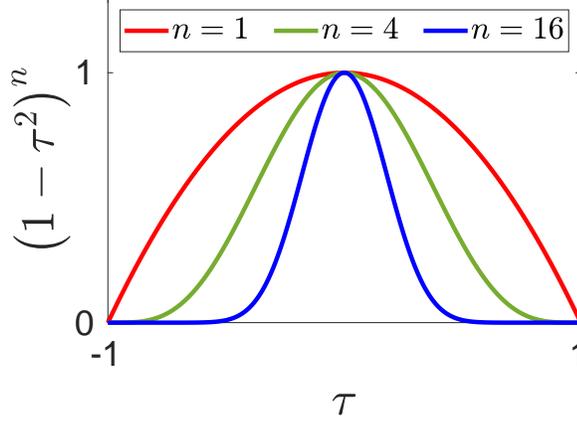


Figure 3.1: **Polynomial envelopes for representative choices of  $n$ .** The higher the value of  $n$ , the fewer grid points are effectively used in averaging.

where  $dV$  is the volume element of  $\Omega_i$ ,  $w_i$  is a weight function, and  $S_j$  is a scale associated with  $\mathbf{f}_j$ . Discussion of the scales  $S_j$  will be delayed until a later section. The domains  $\Omega_i$  are usually taken to be randomly sampled rectangular regions. The rectangular shape of  $\Omega_i$  is natural for gridded data, but this restriction is not essential for the algorithm. On such domains, it is advantageous to use weight functions

$$w_i(t, x, y) = W(\hat{t})W(\hat{x})W(\hat{y}) \quad \hat{x}, \hat{y}, \hat{t} \in [-1, 1], \quad (3.3)$$

$$W(\tau) = (1 - \tau^2)^n. \quad (3.4)$$

The coordinates of  $\hat{\mathbf{x}}$  have been rescaled and shifted to lie on the canonical interval  $[-1, 1]$ . Figure 3.1 shows the shape of  $W(\tau)$  for a few choices of  $n$ . These weight functions and their first  $n - 1$  derivatives vanish on the boundaries. This means integration by parts can be used without introducing boundary terms, allowing numerical differentiation to be avoided in many cases. Derivatives moved onto  $w_i$  can be taken analytically since  $w_i$  is a polynomial. After transferring derivatives, the integral can be evaluated from gridded data using the trapezoidal rule.

Sometimes it is useful to add a modulating factor to the weight function like Legendre polynomials or Fourier modes [34, 37]. These modular factors increase sensitivity to vari-

ation in the data and using multiple weight functions increases the number of independent measurements in  $G$ . This reduces the number of distinct domains that need to be sampled. The product maintains the nice boundary properties of the envelopes: integration by parts is still trivial.

If the functions  $f_j$  are vectors or tensors with multiple components, each component can be taken to be its own row of the matrix. In some scenarios, it is beneficial to consider  $w_i$  to be a tensor as well, so that  $G$  can be defined via an inner product.

$$G_{ij} = \frac{1}{N_i S_j} \int_{\Omega_i} dV \sum_{\alpha} w_{\alpha i} f_{\alpha j} \quad (3.5)$$

where  $\alpha$  is an index over the components of  $f_j$ . This was done in [36] to remove the effects of an latent pressure field.

Integration by parts cannot remove all derivatives from physical fields in general. However, there are two extremely common scenarios where integration by parts can nonetheless be beneficial. The first concerns a product  $f \partial_x g$  where  $f$  and  $g$  are distinct fields. The weighted integral can be integrated by parts as

$$\int_{\Omega} dx w f \partial_x g = - \int_{\Omega} dx w \partial_x f g - \int_{\Omega} dx \partial_x w f g \quad (3.6)$$

where  $w$  is assumed to vanish at the boundary. This manipulation still requires a derivative on  $f$ , but this can be advantageous if  $f$  has been measured with more precision than  $g$ . The second scenario is when the integral is of a product of terms, each with some number  $n_i$  of derivatives acting on fields  $f_i$ . Let  $j$  be the index of the maximal  $n$  such that  $n_j \geq n_i$ . If  $n_j > n_i + 1$  for all  $i \neq j$ , then it is beneficial to move one of these derivatives onto the

other fields<sup>3</sup>.

$$\int_{\Omega} dx w \prod_i \partial_x^{n_i} f_i = \int_{\Omega} dx w \partial_x^{n_j} f_j w \prod_{i \neq j} \partial_x^{n_i} f_i \quad (3.7)$$

$$= - \int_{\Omega} dx \partial_x w \partial_x^{n_j-1} f_j \prod_{i \neq j} \partial_x^{n_i} f_i - \int_{\Omega} dx w \partial_x^{n_j-1} f_j \partial_x \prod_{i \neq j} \partial_x^{n_i} f_i \quad (3.8)$$

Note that the highest power derivative appearing in the integral has been lowered by 1. The last note on manipulating derivatives is that the product rule can sometimes simplify integration. For example,

$$\int_{\Omega} dx w f \partial_x f = \frac{1}{2} \int_{\Omega} dx w \partial_x f^2 = -\frac{1}{2} \int_{\Omega} dx \partial_x w f^2. \quad (3.9)$$

A combination of these tricks can be used to numerically integrate with amazing accuracy. While these have been presented in one dimension for simplicity, they extend to arbitrary dimension.

### 3.2 Introduction to Regression

Once this matrix  $G$  is constructed from the data, many sparse regression techniques can be applied to arrive at parsimonious models. There are two approaches to regression that deserve separate attention.

- **The inhomogeneous problem.** One of the elements of the library  $\mathcal{L}$  is deemed special (usually the time derivative of some quantity). Its associated column  $\mathbf{g}$  of  $G$  is separated. One then seeks an approximate sparse solution  $\mathbf{c}$  to  $G\mathbf{c} = \mathbf{g}$  where  $G$  is now missing the column  $\mathbf{g}$ .
- **The homogeneous problem** All elements of the library  $\mathcal{L}$  are placed on equal footing. No assumptions are made about the presence of any terms. An arbitrary sparse

---

<sup>3</sup>Daniel Gurevich's weak-form implementation [57] automatically detects this case and does the appropriate integration by parts.

linear combination of library terms must approximately vanish. This is equivalent to finding a sparse vector  $\mathbf{c}$  that approximately solves  $G\mathbf{c} = 0$ .

The inhomogeneous problem is easier to tackle since there is an unambiguous measure of success: can you reproduce the right hand side  $\mathbf{g}$ ? One can define the relative residual  $r(\mathbf{c})$ .

$$r(\mathbf{c}) = \frac{\|G\mathbf{c} - \mathbf{g}\|_2}{\|\mathbf{g}\|_2} \quad (3.10)$$

There is a dense vector  $\mathbf{c}^*$  which uniquely minimizes  $r(\mathbf{c})$  assuming that  $G$  is full rank: the least-squares solution obtained from  $G^T G \mathbf{c}^* = G^T \mathbf{g}$ . The error of any sparse relation must be higher than  $r(\mathbf{c}^*)$ , so one should check that  $r(\mathbf{c}^*)$  is sufficiently small. A large minimal residual suggests that the library is incomplete; a more general library should be constructed until the residual is satisfactory.

The homogeneous problem has no direct generalization of  $r(\mathbf{c})$ . For homogeneous problems, the scale of  $\mathbf{c}$  is arbitrary. We fix  $\|\mathbf{c}\|_2 = 1$  for all definitions of residuals, but we usually use a different natural normalization (i.e. set the largest magnitude component to 1) when stating a relation in symbolic form. A natural residual for homogeneous problems is the absolute residual  $\|G\mathbf{c}\|_2$ . There is an exact minimum of  $\|G\mathbf{c}\|_2$  given by  $\mathbf{c}^*$ , which is the right singular vector associated with the smallest singular value of  $G$ . A difficulty with this residual is that it strongly depends on the scaling of each library term (nondimensionalization). An intelligent nondimensionalization scheme should be used when minimizing  $\|G\mathbf{c}\|_2$ . A second useful residual is  $\eta(\mathbf{c})$  [36], which is normalized by the largest term in the relation.

$$\eta(\mathbf{c}) = \frac{\|G\mathbf{c}\|_2}{\max_n \{ \|c_n \mathbf{g}_n\|_2 \}} \quad (3.11)$$

where  $\mathbf{g}_n$  is the  $n$ th column of  $G$ . This residual is bounded from above by the  $\ell_0$  norm of  $\mathbf{c}$ , which helps interpret the accuracy of a homogeneous relation. Single-term models have  $\eta(\mathbf{c}) = 1$  always;  $\eta$  is only meaningful for multi-term relations.

These two residuals  $\|G\mathbf{c}\|_2$  and  $\eta(\mathbf{c})$  are both important considerations for homoge-

neous regression. A low value of  $\|G\mathbf{c}\|_2$  implies quantitative accuracy assuming (i) the data has been properly nondimensionalized and (ii) there are no small columns of  $G$ . Small values of  $\eta(\mathbf{c})$  are a nonnegotiable physical requirement that imply a dominant balance. The absolute residual  $\|G\mathbf{c}\|_2$  can be minimized exactly via the SVD, while  $\eta(\mathbf{c})$  is non-convex and difficult to globally minimize. The two costs functions should be used in parallel to make judgements during model discovery.

To test these various sparse regression schemes in the following sections, we will consider a library matrix of the nematic-invariant rank-2 symmetric trace-free tensors of an active nematic suspension [37] given by Equation 4.23. This library is interesting because it is computed from experimental data and it contains multiple multi-term relations. The details of this library and its evaluation can be found in section 4. One of the library terms is  $\partial_t Q_{ij}$ , which we will use to investigate inhomogeneous algorithms. The lower bound of the inhomogeneous residual of the dynamics is  $r(\mathbf{c}^*) = 0.11$ .

### 3.3 Basis Pursuit Denoising and LASSO

Basis Pursuit Denoising (BPDN) is the optimization problem of finding an approximate solution to a system of equations that is regularized by a  $\ell_1$  penalty [58]. BPDN is formally equivalent to the original formulation of LASSO regression [59], which aims to minimize the cost function

$$f(\mathbf{c}) = \frac{1}{2}\|G\mathbf{c} - \mathbf{g}\|_2^2 + \lambda\|\mathbf{c}\|_1. \quad (3.12)$$

$\lambda \in \mathbb{R}^+$  is the regularization hyperparameter, where larger values  $\lambda$  encourage more aggressive sparsification. The first term in  $f$  penalizes error in  $G\mathbf{c} = \mathbf{g}$ , and the second term penalizes coefficient magnitude. Minimizing this cost function with respect to  $\mathbf{c}$  results in a nonlinear system of equations

$$\frac{\partial f}{\partial \mathbf{c}} = G^T(G\mathbf{c} - \mathbf{g}) + \lambda \text{sign}(\mathbf{c}) = 0, \quad (3.13)$$

which can be solved by a simple discrete search with enough computation time.  $\text{sign}(\mathbf{c})$  has only  $3^{\dim(\mathbf{c})}$  possible values since each element is in  $\{-1, 0, 1\}$ . For low dimensional spaces, one can numerically sweep over all possibilities, solving Equation 3.13 for  $\mathbf{c}$  and checking that  $\text{sign}(\mathbf{c})$  matches the assumption. This exponential scaling makes this exhaustive algorithm impractical for large libraries, but a naive serial MATLAB script can search over the  $3^{13}$  cases in 211 seconds. Many efficient algorithms for BPDN have been proposed in the literature such as the in-crowd algorithm [60], fixed point continuation [61], and homotopy continuation [62] among others.

Table 3.1 shows the results of this sweep for some choices of  $\lambda$ . Sparsification can be quantitatively justified, as  $\lambda = 10^{-1}$  only increases the relative residual by 10% compared to the least squares solution  $\mathbf{c}^*$ . While  $\lambda = 1$  doubles the relative residual compared to  $\mathbf{c}^*$ , it produces a four term right hand side

$$\partial_t Q_{ij} = \hat{P}_{STF} [c_1 u_k \nabla_k Q_{ij} + c_2 \Omega_{ik} Q_{kj} + c_3 A_{ij} + c_4 Q_{ij}] \quad (3.14)$$

with 25% error. An undesirable side effect of  $\ell_1$  regularization is that even if the correct subspace has been identified, the values of coefficients will be biased to minimize the overall cost while increasing the unregularized residual  $\|G\mathbf{c} - \mathbf{g}\|_2$ . For this reason, BPDN should be used to pick a subspace, but not the values of coefficients themselves. A modified coefficient vector  $\mathbf{c}^+$  can be obtained by solving the least squares problem constrained to the subspace identified by LASSO.

Table 3.1: **Characteristics of some exact solutions to BPDN.**  $\lambda$  is the regularization parameter,  $\|\mathbf{c}\|_0$  is the number of nonzero terms in  $\mathbf{c}$ ,  $r$  is the relative residual, and  $\mathbf{c}^+$  is the coefficient vector obtained from the least-squares problem in the same subspace as  $\mathbf{c}$ .

$\lambda$	$\ \mathbf{c}\ _0$	$r(\mathbf{c})$	$r(\mathbf{c}^+)$
$10^{-2}$	11	0.11	0.11
$10^{-1}$	7	0.12	0.11
1	4	0.25	0.16

For  $\lambda = 1$ , the difference between  $\mathbf{c}$  and  $\mathbf{c}^+$  is non-negligible.  $r(\mathbf{c}^+)$  is considerably

lower than  $r(\mathbf{c})$  at no cost to sparsity. The advection coefficient  $c_1$  is expected to be -1 a priori, and  $c_1^+ = -0.99$  is much closer to this compared to  $c_1 = -0.84$ . This is a systematic failure of LASSO in its original form.  $\lambda$  is responsible for both variable selection and magnitude regularization of coefficients. The LASSO method typically struggles when the actual count of significant coefficients is small compared to the total number of coefficients. In such cases, the Lasso faces a dilemma: either include irrelevant variables along with the important ones or excessively shrink the variables, resulting in a model that is appropriately sized but lacks accuracy [63]. LASSO will certainly fail as a proxy for solving the  $\ell_0$  regularization problem for large values of  $\lambda$ , although LASSO can still be useful for subspace selection.

### 3.4 Sequentially Thresholded Least Squares

In the original proposal of SINDy, LASSO was not used because it is expensive for very large data sets in strong form [23]. Instead, the computationally cheap Sequentially Thresholded Least Squares (STLS) algorithm was proposed. This finds a sparse least squares solution to the inhomogeneous problem  $A\mathbf{x} = \mathbf{b}$  by computing a sequence of nested subspaces  $I_{n+1} \subset I_n$ . At each iteration,  $\mathbf{x}_n$  is found as the least squares solution to  $\|A\mathbf{x}_n - \mathbf{b}\|_2$  subject to  $\mathbf{x}_n \in I_n$ . The next subspace is found by hard thresholding the components of  $\mathbf{x}_n$  with magnitude smaller than a parameter  $\lambda$ . This algorithm can be written in only a few lines of MATLAB.

```
function x = STLS(A,b,lambda)
    %Modified from Brunton 2016
    x = A \ b; %initial guess: Least squares
    for k = 1:10
        big = abs(x) > lambda; %find large coefficients
        x(not(big)) = 0; %set small values to 0
        x(big) = A(:,big) \ b; %redo least-squares
    end
end
```

end

end

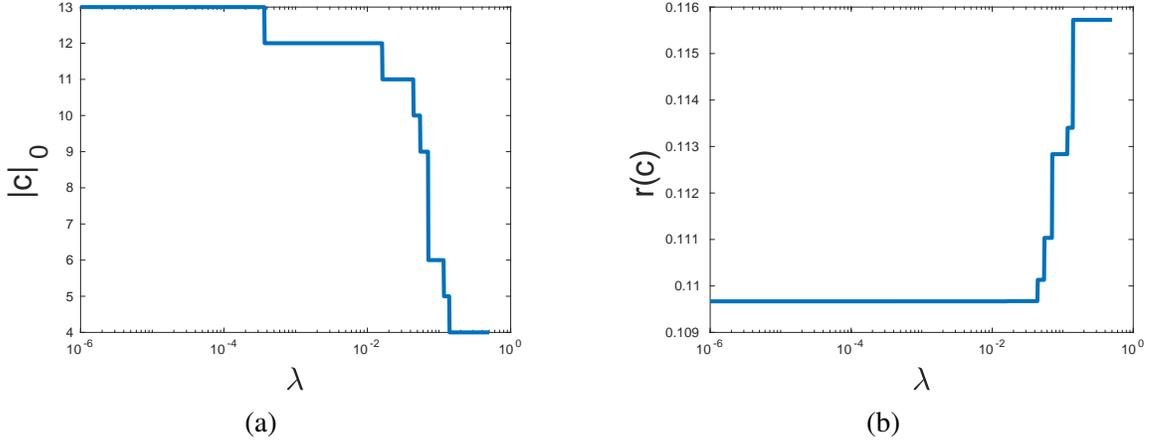


Figure 3.2: **STLS applied to the rank-2 library of active nematic data. (a) shows the number of nonzero terms in  $c$  as a function of  $\lambda$ . (b) shows the relative residual  $r(c)$  as a function of  $\lambda$ .**

While the algorithm should halt in finitely many iterations when  $\mathbf{x}$  is a fixed point, a hard cutoff of 10 iterations is usually sufficient for  $\mathbf{x}$  to converge. This algorithm is incredibly fast, so it is possible to do a thorough sweep over  $\lambda$  for the active nematics library. Such a sweep is illustrated in Figure 3.2 with  $10^{-6} < \lambda < 0.5$ . As seen in Figure 3.2b, the residual  $r(c)$  remains close to optimal, even up to  $\lambda = 0.5$ . There are only 4 nonzero terms in this parameter regime, so we take this as the optimal STLS model.

$$\partial_i Q_{ij} = \hat{P}_{STF} [c_1 u_k \nabla_k Q_{ij} + c_2 \Omega_{ik} Q_{kj} + c_3 A_{kl} Q_{kl} Q_{ij} + c_4 A_{ij}] \quad (3.15)$$

The coefficients are  $O(1)$ , and increasing  $\lambda$  to unity quickly leads to  $c = 0$ . **Curiously, the sparse models discovered by LASSO and STLS do not agree.** Equation 3.14 is missing  $A_{kl} Q_{kl} Q_{ij}$  and is using the simpler term  $Q_{ij}$  instead. This discrepancy motivates the next section.

While STLS outperforms LASSO for the active nematic library, it does face difficulties when small coefficients are present. Even if a small coefficient significantly improves the

residual, thresholding will remove it. Small coefficients are not uncommon in diffusive problems. To demonstrate this, one can integrate a library containing the Navier-Stokes equation with high Reynolds number channelflow data from the Johns Hopkins Turbulence database.

$$\partial_t u_i = c_1 \nabla_j (u_j u_i) + c_2 \nabla_i p + c_3 \nabla^2 u_i \quad (3.16)$$

A least squares solution  $\mathbf{c}^*$  for a library containing 12 terms results in  $c_3 = 5.3 \times 10^{-5}$  after nondimensionalization, which is orders of magnitude smaller than coefficients of spurious terms in the library. STLS will never be able to recover the correct governing equation because of the small parameter problem.

### 3.5 Greedy Iterative SVD Regression

If one makes no assumptions about which terms exist in a given empirical relation, then the problem at hand is the homogeneous one: find a sparse vector  $\mathbf{c}$  such that  $G\mathbf{c} \approx 0$ . The error can be minimized by taking  $\mathbf{c}$  to be the smallest right singular vector  $\mathbf{c}^*$  of  $G$ , that is the right singular vector associated with the smallest singular value.  $\mathbf{c}^*$  is generally dense, but it is guaranteed to minimize the  $\ell_2$  norm of the residual.

We propose a greedy<sup>4</sup> algorithm for obtaining sparse solutions to  $G\mathbf{c} \approx 0$ , which utilizes a sequence of nested sparse subspaces. Only one coefficient is set to zero at a time, and a coefficient cannot become nonzero after it has been discarded. Since  $\mathbf{c}^*$  minimizes the residual  $\|G\mathbf{c}\|_2$  exactly, enforcing vanishing coefficients will only increase this residual. It follows that any iterative sparsification procedure utilizing nested sparse subspaces must produce a monotonically increasing residual as the number of nonzero terms goes down. Greedy Iterative SVD Regression (GISR) iteratively removes the least informative library term and recomputes the right singular vector after each term is discarded. A term is “least informative” if its removal results in the least increase in residual. The algorithm is halted when the residual increases by a multiplicative factor  $\gamma$ , which we usually take to be some

---

<sup>4</sup>Greedy algorithms choose the best immediate option without consideration of the future.

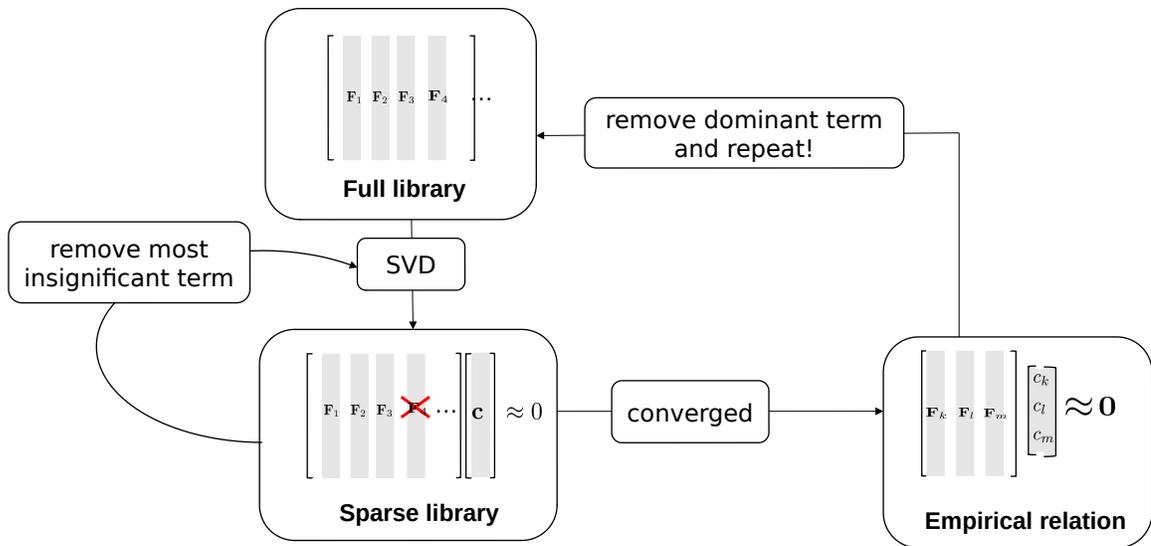


Figure 3.3: **A schematic of GISR.** Terms are greedily removed from the library to obtain a sparse approximate null vector. The SVD is recomputed after every modification of the library.

value in the interval  $[1.05, 1.5]$ .

```
function [C, n] = GISR( G, gamma )
    m = size(G,2); %number of right hand side
    idx = 1:m;    %indices of used terms
    C = zeros(m,m); %columns correspond to models

    for i = m:-1:1
        [~,~,V] = svd(G(:,idx)); %compute right singular vectors
        C(idx,i) = V(:,end);    %save last vector

        if i == 1
            break;
        end

        useless = 0; %index of the most "useless" term
```

```

min_res = inf; %keep track of the best residual
for j = 1:numel(idx)
    idx_temp    = idx;
    idx_temp(j) = [];
    s = svd( G(:, idx_temp) );
    if( s(end) < min_res )
        min_res = s(end);
        useless = j;
    end
end
idx(useless) = [];
end
r = vecnorm(G*C);
n = 1 + find( r(1:m-1) ./ r(2:m) > gamma , 1, 'last' );
end

```

GISR uses multiplicative increase in residual as a halting condition. This method is not well-suited for single term models like Gauss's law  $\nabla \cdot \mathbf{B} = 0$ . While these single term models can in principle be found by halting when a single term is reached, this is grossly computationally inefficient. Instead, single term models can be identified by evaluating the norms of columns and comparing them to some cutoff, e.g. if a column  $\mathbf{g}$  has a norm  $\|\mathbf{g}\|_2 < 10^{-3}\sqrt{m}$ , where  $m$  is the number of observations composing  $\mathbf{g}$ . This comparison is only possible because of nondimensionalization. For noisy data, a higher threshold might be needed to identify single term relations. Regardless of the success of this heuristic identification, GISR should still halt on single term models provided they are properly nondimensionalized.

Multiple independent relations can exist within the same library matrix  $G$ , so it is useful to have a method capable of finding them. This can be accomplished by permanently

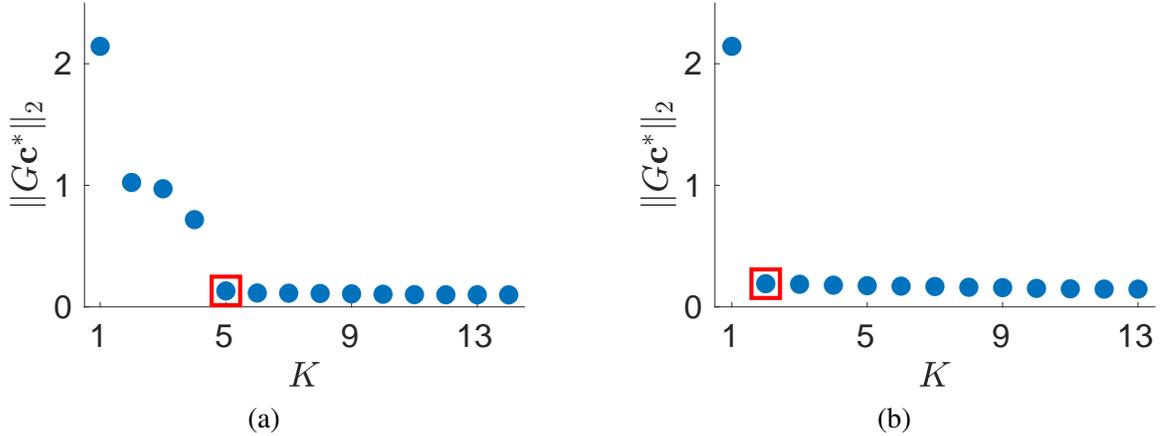


Figure 3.4: **Residual curves for two subsequent runs of GISR applied to a rank-2 library.** The first run finds a five term relation. After the largest term is discarded from the library, a subsequent run finds a 2-term relation.

removing the largest term of an identified relation from the library, and repeating GISR. A useful definition of largest term is  $\max_n \|c_n \mathbf{g}_n\|_2$ , where  $\mathbf{g}_n$  is the  $n$ th column of  $G$ . This can be repeated indefinitely until GISR no longer returns sparse relations with low residuals. The results of GISR applied to the rank-2 active nematics library can be seen in Figure 3.4.

### 3.6 Reverse Regression: a holistic approach to sparse regression

A weakness of GISR is the high number of SVD calculations it requires. Figure 3.5 shows the computation time needed for GISR on a library matrix of size  $2n \times n$ . This size is chosen so that  $G$  is overdetermined, otherwise there will be a spurious right null space that will immediately halt the algorithm. For modest libraries of size  $n = 256$  or less, the  $n^4$  scaling is manageable. A larger issue is the difficulty of resolving small singular values and their associated vectors for ill-conditioned matrices. If the dense singular vector associated with smallest singular value cannot be well resolved, then the earliest steps of greedy sparsification cannot be trusted. These difficulties motivate Reverse Regression (RR). RR consists of four steps.

1. Instead of initially considering the entire library (“the whole kitchen sink”), recog-

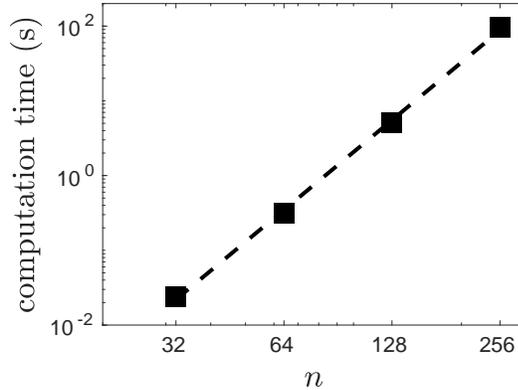


Figure 3.5: **Scaling of the GISR computation time with library size.** The dashed line shows a polynomial fit with scaling  $n^4$ .

nize that a general multi-term relation will have a dominant balance. A small number of terms, perhaps just two, will have a high correlation. One can do a cheap combinatoric search of all pairs of terms to identify a dominant balance. Genetic algorithms could also be useful in determining an initial dominant balance since the number of terms is constrained. Alternatively, if one knows some of the physics a priori (possibly when debugging numerical simulations), a many-term initial guess can be taken.

2. The reverse of GISR is performed. That is, terms are **added** to the relation greedily to decrease the residual as much as possible. The halting condition for this step is nontrivial. As seen in Figure 3.7, insignificant improvement is achieved in going from  $N = 3$  to  $N = 8$ , but then rapid improvement occurs in the next several terms. One should halt adding terms when the residual has not meaningfully decreased for  $O(10)$  steps.
3. Once the residual has stopped decreasing, GISR can be applied to the currently included terms to remove any spurious terms added in step 2. Sparsification is halted when the residual increases by some factor  $\gamma \in [1.05, 1.5]$ .
4. Once a sparse relation is obtained, the dominant term can be permanently removed from the library and RR restarted from the sublibrary. An arbitrary number of relations can be identified in this fashion.

As an example of RR, let us consider the lambda-omega reaction diffusion equations describing two chemical concentrations  $u$  and  $v$ .

$$\partial_t u = (1 - s^2)u + \beta s^2 v + \sigma \nabla^2 u \quad (3.17)$$

$$\partial_t v = (1 - s^2)v - \beta s^2 u + \sigma \nabla^2 v \quad (3.18)$$

where  $s^2 = u^2 + v^2$ . Numerical data is generated on a  $256 \times 256 \times 512$  grid with RK6 and parameters  $\beta = 3$  and  $\sigma = 0.1$ . The boundary conditions are  $\hat{n} \cdot \nabla u = \hat{n} \cdot \nabla v = 0$ . The space and time coordinates are in the intervals  $x, y \in [0, 1]$  and  $t \in [0, 5.12]$ . Figure 3.6 shows the final state of integration.

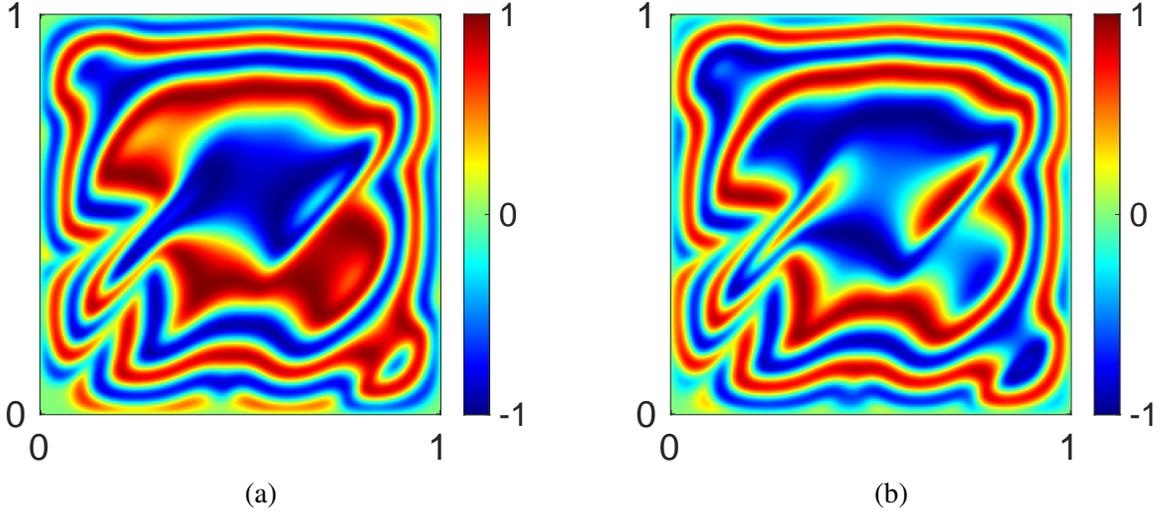


Figure 3.6: **Final state after integration of the lambda-omega reaction diffusion equations.** (a) shows  $u$  and (b) shows  $v$ .

If no symmetry is assumed, one can expand in polynomial powers to obtain a 21-term library.

$$\mathcal{L} = \{ \partial_t u, \partial_t v, u, v, u^2, uv, v^2, u^3, u^2 v, uv^2, v^3, \partial_x u, \partial_x^2 u, \partial_y u, \partial_y^2 u, \partial_x \partial_y u, \partial_x v, \partial_x^2 v, \partial_y v, \partial_y^2 v, \partial_x \partial_y v \} \quad (3.19)$$

$G$  is constructed to be of size  $42 \times 21$  with 6th power weight functions and  $64 \times 64 \times$

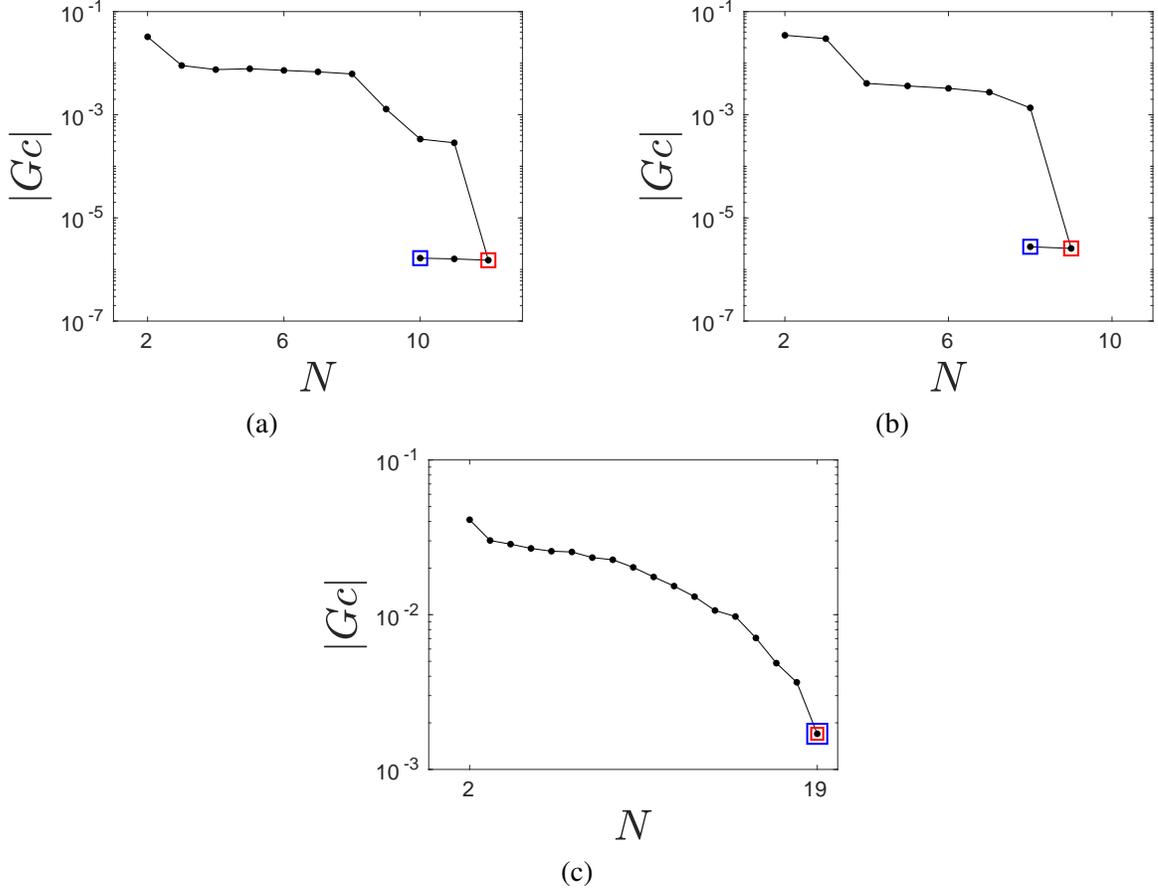


Figure 3.7: **Reverse regression curves of the reaction-diffusion library starting from 2-term dominant balances.** The red marker indicates where no more terms need added. The blue marker indicates where no more terms can be removed. (a), (b), and (c) start RR with Equation 3.20, Equation 3.22, and Equation 3.24, respectively.

64 gridpoints used in each subdomain  $\Omega_i$ . The only weight function considered on each subdomain is the product of polynomial weights  $w(\tau) = (1 - \tau^2)^6$  in each dimension. A combinatoric search reveals that the best dominant balance is

$$\partial_t v + 4u^3 \approx 0 \quad (3.20)$$

with a relative residual of 0.15. Figure 3.7 shows that RR initially finds a 12 term relation

and then is able to remove two spurious terms before halting. The final 10 term relation is

$$\partial_t u + 3\partial_t v - 3v - u + 10u^3 + 10v^2u - 0.1\partial_x^2 u - 0.1\partial_y^2 u - 0.3\partial_x^2 v - 0.3\partial_y^2 v = 0. \quad (3.21)$$

Coefficients are correct to four decimal places. This is not immediately the generating dynamics of Equation 3.18, rather it is a linear combination of  $\partial_t u$  and  $\partial_t v$  that annihilates  $v^3$  and  $vu^2$ . If multiple relations exist in a library, sparse regression often leads to linear combinations of those relations, especially if this linear combination has a lower value of  $\|Gc\|_2$ . The largest term of this relation is  $\partial_t v$ , which can be removed from the library permanently and RR repeated. The new best dominant balance captures the leading dynamics of  $u$ .

$$\partial_t u - 2.8v \approx 0 \quad (3.22)$$

As seen in Figure 3.7(b), adding terms results in a 9 term relation that can have one term removed. Unsurprisingly, the identified relation is none other than the dynamics of  $u$  with the largest term being  $\partial_t u$ . Coefficients are again correct to four decimal places.

$$\partial_t u - u - 3v^3 - 3u^2v + u^3 + uv^2 - 0.1\partial_x^2 u - 0.1\partial_y^2 u = 0 \quad (3.23)$$

After removing both  $\partial_t u$  and  $\partial_t v$  from the library, RR can be performed again to search for a third relation. It is possible other relations exist that capture the geometry of the attractor. The dominant algebraic balance is a polynomial in  $u$ .

$$0.75u - u^3 \approx 0 \quad (3.24)$$

RR does not find a sparse relation from the dominant balance. Adding terms never stops improving the residual as seen in Figure 3.7(c). GISR likewise halts immediately at the dense relation.

### 3.7 Hyperparameter Selection

The success of sparse regression depends on making intelligent choices for weak formulation hyperparameters: weight function parameters and nondimensionalization via the scales  $S_j$ . To illustrate the effects of various choices, we will consider the Lorenz attractor.

$$\dot{x} = \sigma(y - x) \quad \dot{y} = x(\rho - z) - y \quad \dot{z} = xy - \beta z \quad (3.25)$$

with  $\sigma = 10$ ,  $\rho = 28$ , and  $\beta = 8/3$ . Integrating forward in time converges on the famous butterfly attractor. The dataset used for model discovery can be seen in Figure 3.8. As a

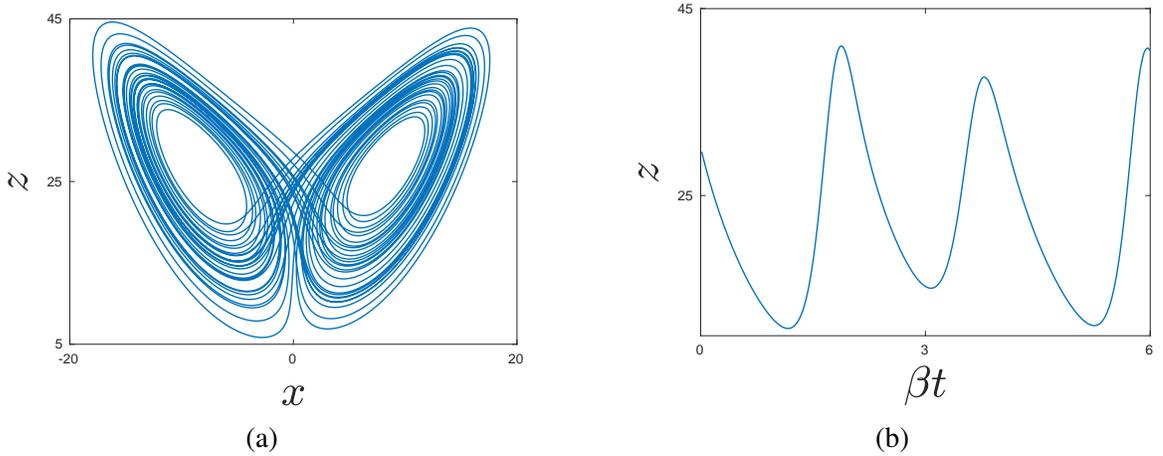


Figure 3.8: **The Lorenz attractor.** (a) A 2D projection of a numerical trajectory generated with RK4 with  $dt = 1/200$ . The initial transient is removed so that the trajectory lies very close to the attractor. (b) Numerical evidence that  $T = \beta^{-1}$  is a reasonable timescale for  $z$  dynamics on the attractor.

test of hyperparameter choices, we use a library containing only the dynamics  $\dot{z}$ . Without appealing to symmetry, a library up to quadratic order can be used.

$$\mathcal{L} = \{\partial_t z, x, y, z, x^2, y^2, z^2, xy, xz, yz\} \quad (3.26)$$

This library will be used to test different nondimensionalization schemes. Here is a short list of potential methods.

1. **Ostrich nondimensionalization (OND):** The trivial choice is to do nothing. Take  $S_j = 1$  so that nondimensionalization is skipped and no thinking is required. This cannot be recommended but represents a useful benchmark.
2. **Unity nondimensionalization (UND):** This physics-free scheme chooses  $S_j$  so that the norm of each column of  $G$  is unity.
3. **Simple nondimensionalization (SND):** Nondimensionalize the fundamental variables and coordinates with their means.  $S_j$  are defined implicitly by rescaling the variables from the data before computing any function  $f_j \in \mathcal{L}$ .

$$x' = \frac{x}{\langle x \rangle} \quad y' = \frac{y}{\langle y \rangle} \quad z' = \frac{z}{\langle z \rangle} \quad (3.27)$$

$$t' = \frac{t}{T} \quad (3.28)$$

There are many choices for the timescale  $T$  which are sensible. One can either choose  $T$  to be the timescale of a known physical process or use it to fix the velocity scale  $\langle |\dot{x}'| \rangle$  to be unity. For the Lorenz system, one can take  $T = \beta^{-1}$  since  $\beta$  is a decay timescale of the Lorenz system. This scheme was used in [37] by choosing the length and time scales which set the mean flow and vorticity to unity:  $\langle |\mathbf{u}| \rangle = \langle |\nabla \times \mathbf{u}| \rangle = 1$ .

4. **Physical nondimensionalization (PND):** Each scale  $S_j \equiv S[f_j]$  can be determined independently by writing  $f_j$  as a product of simple terms. Here the scales of terms are uniquely determined by the rules  $S[fg] = S[f]S[g]$ ,  $S[v] = \mu[v]$ , and  $S[\partial_b^n v] = B^{-n}\sigma[v]$ . Here  $v$  is a fundamental observable of the system and  $B$  is a characteristic scale associated with the spacetime coordinate  $b$ .  $\mu[v]$  and  $\sigma[v]$  are the mean and standard deviation of  $v$ , respectively. Different time and length scales can be chosen for each (possibly indexed) field  $\mathbf{f}$  if desired.

$$T_{\mathbf{f}} = \frac{\sigma[\mathbf{f}]}{\mu[\sqrt{\partial_t \mathbf{f}} \cdot \partial_t \mathbf{f}]} \quad L_{\mathbf{f}} = \frac{\sigma[\mathbf{f}]}{\mu[\sqrt{\nabla \mathbf{f}} : \nabla \mathbf{f}]} \quad (3.29)$$

To test the efficacy of these schemes, we examine the likelihood of discovering the correct  $\dot{z}$  with GISR from the library of Equation 3.26. We can measure the probability of discovering the correct relation as a function of noise and the number of subdomains (observations)  $d$ . Here each variable  $x_i$  is contaminated with uniform random noise:

$$x_i^{\text{noisy}} = x_i + \sigma s_i N \quad (3.30)$$

where  $\sigma$  is the noise level,  $s_i$  is the standard deviation of each coordinate  $x_i$ , and  $N$  is uniform random noise from  $[-1, 1]$ . Noise is added after the estimation of  $S_j$ .

Figure 3.9 shows that the choice of scheme is important. Failing to nondimensionalize at all via the OND scheme results in very poor noise robustness. Noise levels  $\sigma \gtrsim 0.1$  destroy any chance of successful identification of the dynamics. The unphysical UND scheme is an improvement, allowing noise levels up to  $\sigma \approx 0.25$ . The physics-informed nondimensionalizations SND and PND offer far superior robustness, with SND having high success rates in all cases. This figure also demonstrates a weak dependence on the number of sampled subdomains  $d$  once the system has been overdetermined.

Another hyperparameter that can be tuned is the envelope exponent  $n$  of Equation 3.1.  $n$  should be larger than the highest order derivative in the library to ensure success of integration by parts. For data on a uniform grid, the error in coefficients (stemming from error in quadrature) scales like  $h^{-n}$  where  $h$  is the uniform grid spacing [25, 34]. This scaling suggests the higher the power, the better. To investigate the quadrature error of the trapezoid rule, consider the following analytic test case integrating  $\cos(2\pi\tau)$ .

$$I_n = \int_{-1}^1 d\tau (1 - \tau^2)^n \cos(2\pi\tau) = \frac{\Gamma(n+1) J_{n+\frac{1}{2}}(2\pi)}{\pi^n} \quad (3.31)$$

$$r_n \equiv \frac{|I_n - T_n|}{|I_n|} \quad (3.32)$$

where  $T_n$  is the trapezoidal rule approximation of  $I_n$ ,  $\Gamma(n)$  is Euler's Gamma function, and

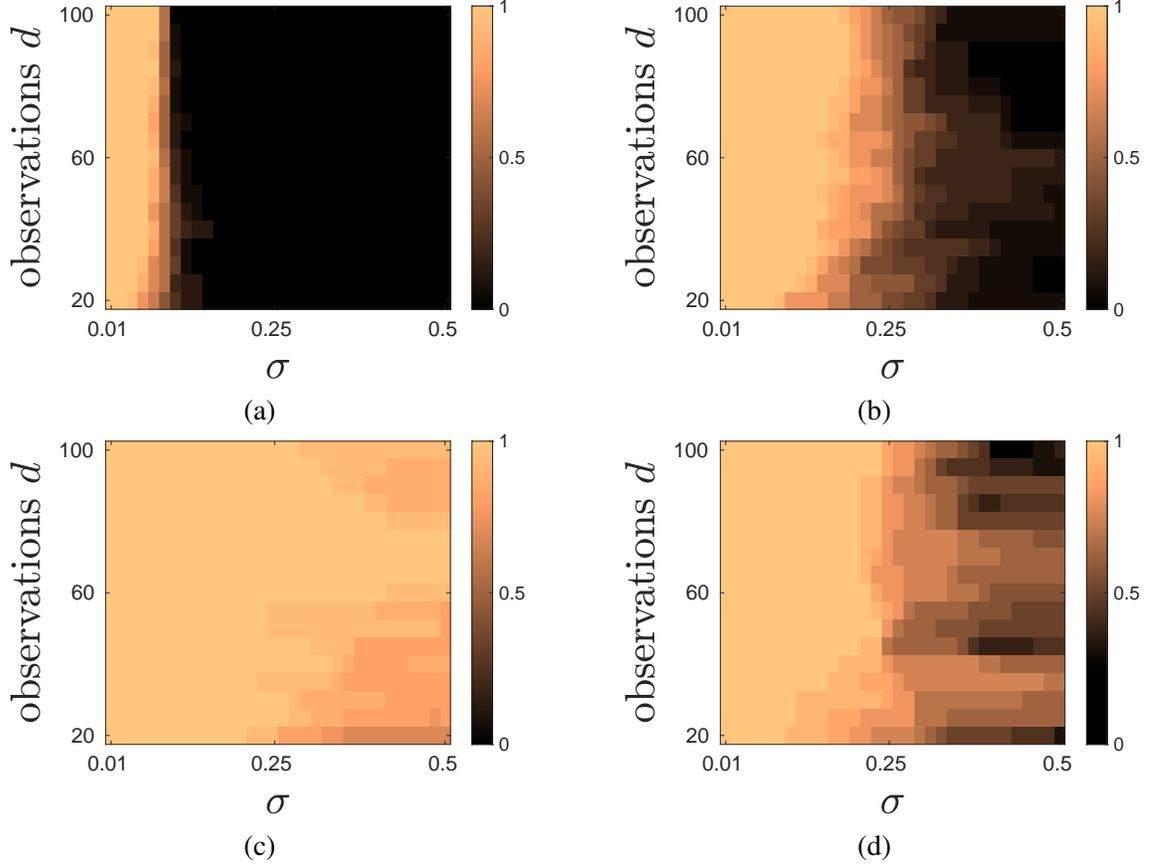


Figure 3.9: **Probability of discovering the correct dynamics of  $z$  with various nondimensionalization schemes, noise levels, and number of domains.** Integrals for the weak form are computed with 256 points in time and envelope power 6. This corresponds to a nondimensional window size of  $\beta t_\Omega \approx 3.4$ . Since the library contains 10 terms, each matrix  $G$  is of size  $d \times 10$ . To estimate the correct discovery probability, subdomains are chosen randomly 16 times and GISR is performed independently for each sampling with  $\gamma = 1.5$ . (a) OND. (b) UND. (c) SND. (d) PND.

$J_n(x)$  is a Bessel function of the first kind. Let  $x_n$  be 32 points on the interval  $[-1, 1]$  spaced uniformly and  $y_n$  be a nonuniform grid corresponding to  $x_n$  with *interior points* randomly shifted by uniform noise of amplitude  $0.01h$ . As shown in Figure 3.10, this small modification in grid location makes the error fail to converge exponentially, and raises the noise floor by orders of magnitude.

There are circumstances where data are not available on a uniform grid without interpolation. For example, turbulent channel flow data from the Johns Hopkins Turbulence Database are simulated on Chebyshev nodes. In these situations, Figure 3.10 should be re-

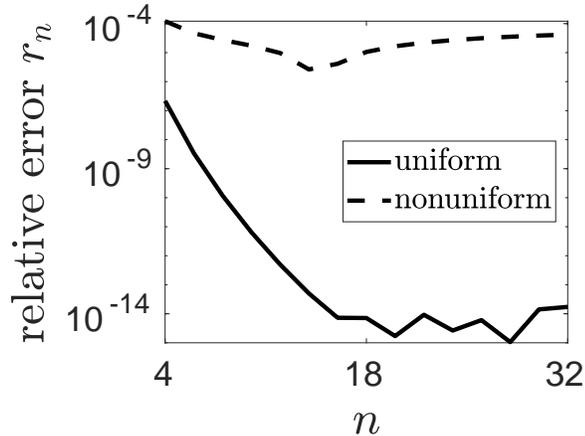


Figure 3.10: **Scaling of the relative residual of the trapezoid rule.** The points on the nonuniform grid have only been moved up by to 1% of the grid spacing.

produced using a nonuniform grid typical of the available data as a guideline for selecting  $n$ .  $n$  can be chosen to minimize the error from quadrature in the uniform direction if such a direction exists.

### 3.8 Data-Dependent Weight Functions

The weight functions considered so far have been polynomials and analytic modulations that have many derivatives vanishing on the boundaries. While such a choice is convenient for rapid integration, there are some situations where a more complex weight function is warranted. If  $w = w_{\text{envelope}}\psi$ , where  $\psi$  is a data-dependent weight function,  $w$  retains the desirable boundary properties of the polynomial factor. Integration by parts becomes more painful since  $\psi$  will need to be numerically differentiated.

The first situation to introduce  $\psi$  is when the data contain singularities or discontinuities. Such a case arose in applying SPIDER to the active nematic suspension [37], where point defects have singularities in the derivatives of the normalized nematic tensor  $Q_{ij}$ . The following MATLAB code demonstrates an algorithm for producing a smooth mask  $\psi$  of unwanted data. This mask is **exactly zero** in regions of unwanted data and has smooth numerical derivatives if sufficiently many iterations are performed.

```

function mask = generate_mask(unwanted)
    mask = ones(size(unwanted)); %initialize the mask
    for k = 1:10
        mask = imgaussfilt(mask,2); %blur the mask
        mask(unwanted) = 0; %force the mask to vanish
        mask = mask/max(mask); %normalize
    end
    mask = mask.^10; %Take a high power to smooth
end

```

The mask  $\psi$  is constructed by repeatedly setting  $\psi$  to zero in regions of unwanted data, smoothing  $\psi$ , and finally taking a high power to smooth the mask near  $\psi = 0$ . Figure 3.11 shows the typical output of this algorithm and evidence that the mask has a well-behaved numerical derivative.

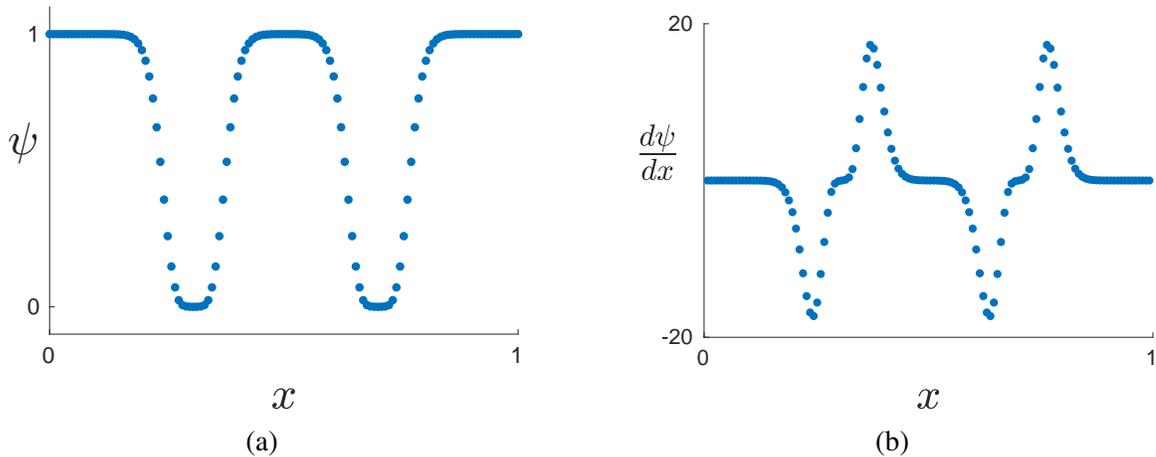


Figure 3.11: **A mask and its derivative computed with the MATLAB function `generate_mask`.** Two points at  $x \approx 0.3$  and  $x \approx 0.7$  have been flagged as ‘bad’ so that  $\psi = 0$  exactly there. (a) shows the constructed mask and (b) shows the derivative estimated with a centered difference. Higher derivatives should be evaluated with regularized difference schemes given the fast variation of  $\psi$ .

Another approach that has been used is to define a positive data-dependent quantity  $\chi$  that diverges and directly compute a mask from it. In the active nematics data, the

determinant of the gradient blows up at topological defects.

$$\chi = |\partial_x n_x \partial_y n_y - \partial_x n_y \partial_y n_x| \quad (3.33)$$

A three parameter mask was constructed for the nematic data using a blurring radius  $\sigma$ , scale  $\alpha$  and power  $\beta$ , where all parameters are  $O(1)$ .  $\mathcal{B}_\sigma$  is a Gaussian convolution with width  $\sigma$ , and  $\langle \cdot \rangle$  denotes an average over available data.

$$\psi = \mathcal{B}_\sigma \left[ \tanh \left( \frac{\alpha \exp(\langle \ln \mathcal{B}_\sigma(\chi) \rangle)}{10^{-9} + \mathcal{B}_\sigma(\chi)} \right)^\beta \right] \quad (3.34)$$

This formula requires only two convolutions and some tuning of algebraic parameters to achieve satisfactory smoothness. Figure 3.12 shows a typical frame of  $Q_{11}$  from experiment and the associated mask  $\psi$  that was constructed to remove the defects. A further difficulty of the nematic symmetry  $\mathbf{n} \equiv -\mathbf{n}$  is that there will be unresolvable discontinuities in  $\mathbf{n}$ . The resolution of these discontinuities in the weak form will be discussed in Chapter 4 (see Figure 4.3). Practically, this means one must be careful when finite differencing the field. Neighboring  $\mathbf{n}$ 's can be multiplied by  $-1$  in the finite difference stencils to minimize the absolute value of the derivative.

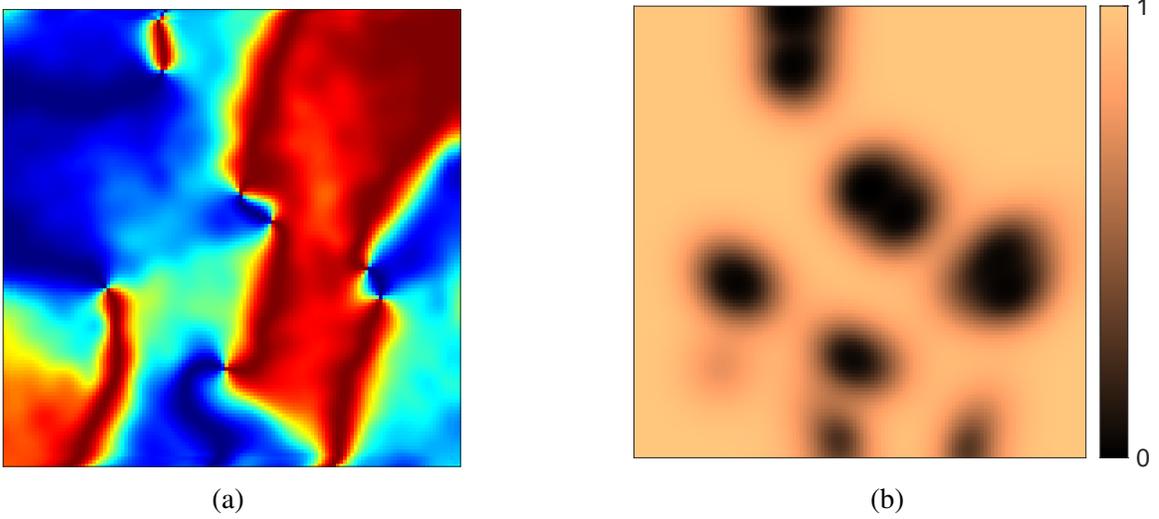


Figure 3.12: **A mask constructed to mask topological defects.** (a) shows a snapshot of  $Q_{11}$ , a component of the nematic tensor  $Q_{ij}$ . The color bar is arbitrary. Topological defects can be identified as points with abrupt change in  $Q_{11}$ . (b) shows a computed  $\psi$  that goes to zero smoothly around these point-like defects.

## CHAPTER 4

### APPLICATIONS OF SPIDER

An immediate application of SPIDER is the validation of numerical simulations. Any numerical discretization, in particular approximations of derivatives, introduces error into PDE solutions. SPIDER can independently estimate derivatives in the weak formulation and validate that data well approximates solutions to a set of governing equations. This will be demonstrated on numerical simulations of 3D fluid turbulence. SPIDER is primarily designed to identify unknown physical relations from experimental data. This application will be demonstrated on an experimental microtubule suspension, which exhibits self-driven dynamics. A complete PDE model is identified by SPIDER using a sequence of 2D images, and some consequences of this model are considered.

#### 4.1 Direct Numerical Simulations of Turbulence

The libraries discussed in Chapter 2 for fluid dynamics systems with  $O(n)$  symmetry can be tested on a numerical solution of the 3D Navier-Stokes equation. We use data of a flow through a rectangular channel from the Johns Hopkins University turbulence database (<http://turbulence.pha.jhu.edu/Channel.Flow.aspx>) [64]. The database provides velocity and pressure data with periodic boundary conditions in two directions and no-slip in the third. The channel dimensions are  $L_x \times L_y \times L_z \times L_t = 8\pi \times 2 \times 3\pi \times 26$  (in nondimensional units) and the fields are stored on a spatiotemporal grid of size  $2048 \times 512 \times 1536 \times 4000$ . The kinematic viscosity is  $\nu = 5 \times 10^{-5}$  and the flow is driven by a constant mean pressure gradient of  $\langle \partial_x p \rangle = 0.0025$ . The pressure can only be defined locally due to this mean gradient, so the database stores the spatially periodic pressure with the mean gradient subtracted. Note that the pressure is not essential as the governing equations can be identified from velocity data alone [36]. Velocity and pressure data was sampled from the middle of

the channel on a  $64^4$  gridpoints. Integration domains are 32 gridpoints on each side. Grid spacing is uniform in every direction but wall-normal, where a Chebyshev nodes are used. 256 subdomains are sampled, and derivatives are calculated with second order finite differences where needed. As discussed in chapter 3 (Figure 3.10), the nonuniform grid spacing changes the convergence of numerical integration. To effectively minimize integration error of a test function, we choose an envelope power of 8. Nondimensionalization is done with PND, where the pressure and velocity are assigned separate length and time scales. The standard deviation of pressure is computed with the periodic pressure.

The scalar library Equation 2.32 contains three single term relations:  $\nabla_i u_i$ ,  $p \nabla_i u_i$ , and  $(\nabla_i u_i)^2$  all have exceptionally small magnitudes. Since these terms have been nondimensionalized, they can be compared to unity. The norms of their columns divided by the square root of the number of samples is  $6 \times 10^{-6}$ ,  $4 \times 10^{-4}$ , and  $6 \times 10^{-4}$ , respectively. This justifies the incompressibility condition  $\nabla_i u_i = 0$ , which can immediately be pruned from all following libraries. Furthermore, if a term can be evaluated nicely by adding or subtracting  $\nabla_i u_i$ , this will be done without comment.

The application of GISR with  $\gamma = 1.5$  to the scalar and vector libraries can be seen in Figure 4.1. GISR first yields a five-term energy relation, and the largest term representing energy advection is pruned. The second application of GISR finds a pressure Poisson equation with a small modification ( $|c_8| \ll 1$ ):

$$c_1 \partial_t E + c_2 \nabla_i (u_i E) + c_3 u_i \nabla_i p + c_4 \nabla^2 E + c_5 (\nabla_i u_j) (\nabla_i u_j) = 0, \quad (4.1)$$

$$c_6 \nabla^2 p + c_7 \nabla_i (u_j \nabla_j u_i) + c_8 = 0, \quad (4.2)$$

where  $E = u_i u_i / 2$  is the kinetic energy. Removing  $\nabla^2 p$  from the library and running GISR a final time produces a phenomenological transport relation for pressure:  $\partial_t p + 0.96 u_i \nabla_i p + 0.002 u^2 = 0$ . The residual of this equation is two orders of magnitude higher than the previous two, indicating that this is not a valid relation in strong form. However, this

relation still offers physical insight into effective sources of pressure.

Applying GISR to the vector library  $\mathcal{L}_1$  identifies a five-term momentum relation.

$$c_9 \partial_t u_i + c_{10} u_j \nabla_j u_i + c_{11} \nabla_i p + c_{12} \nabla^2 u_i = 0 \quad (4.3)$$

Removing  $u_j \nabla_j u_i$  and repeating GISR does not find any physically informative relations.

**Table 4.1: Table of identified coefficients.** The uncertainties  $s_n$  are estimated by the standard deviation of coefficients when regression is done 100 times with different integration subdomains. The magnitude of terms  $\chi_n = \|c_n \mathbf{g}_n\| / \max\{\|c_m \mathbf{g}_m\|\}$  determines how much each term contributes to the relation.

	$\partial_t E$	$\nabla_i(u_i E)$	$u_i \nabla_i p$	$\nabla^2 E$	$(\nabla_i u_j)(\nabla_i u_j)$
$c_n$	0.99339152	0.99321	1	$-4.94 \times 10^{-5}$	$4.851 \times 10^{-5}$
$s_n$	$2 \times 10^{-5}$	$8 \times 10^{-6}$	$5 \times 10^{-4}$	$1 \times 10^{-7}$	$2 \times 10^{-6}$
$\chi_n$	1	1	0.07	0.006	0.001

	$\nabla^2 p$	$\nabla_i(u_j \nabla_j u_i)$	1
$c_n$	1	0.999788	0.00038
$s_n$	$7 \times 10^{-5}$	$5 \times 10^{-6}$	$1 \times 10^{-5}$
$\chi_n$	1	1	0.002

	$\partial_t u_i$	$u_j \nabla_j u_i$	$\nabla_i p$	$\nabla^2 u_i$
$c_n$	0.9998556	0.999858	1	$-5.003 \times 10^{-5}$
$s_n$	$2 \times 10^{-7}$	$1.4 \times 10^{-6}$	$2 \times 10^{-5}$	$2 \times 10^{-8}$
$\chi_n$	0.98	1	0.06	0.006

The rank-2 libraries will not be discussed for this system, but they will be used in the case of active nematics. The values of coefficients are given in Table 4.1. The energy equation has two terms stemming from viscosity: heat-flux  $\nabla^2 E$  and viscous dissipation  $(\nabla_i u_j)(\nabla_i u_j)$ . The recovered viscosity coefficients  $c_4$  and  $c_5$  have 1% and 3% errors, respectively. This is significantly less accurate than  $c_{12}$  with 0.06% error. The reason for this is that all derivatives in the Navier-Stokes equation can be transferred onto the weight function, so numerical differentiation is never used. Unfortunately this is not true

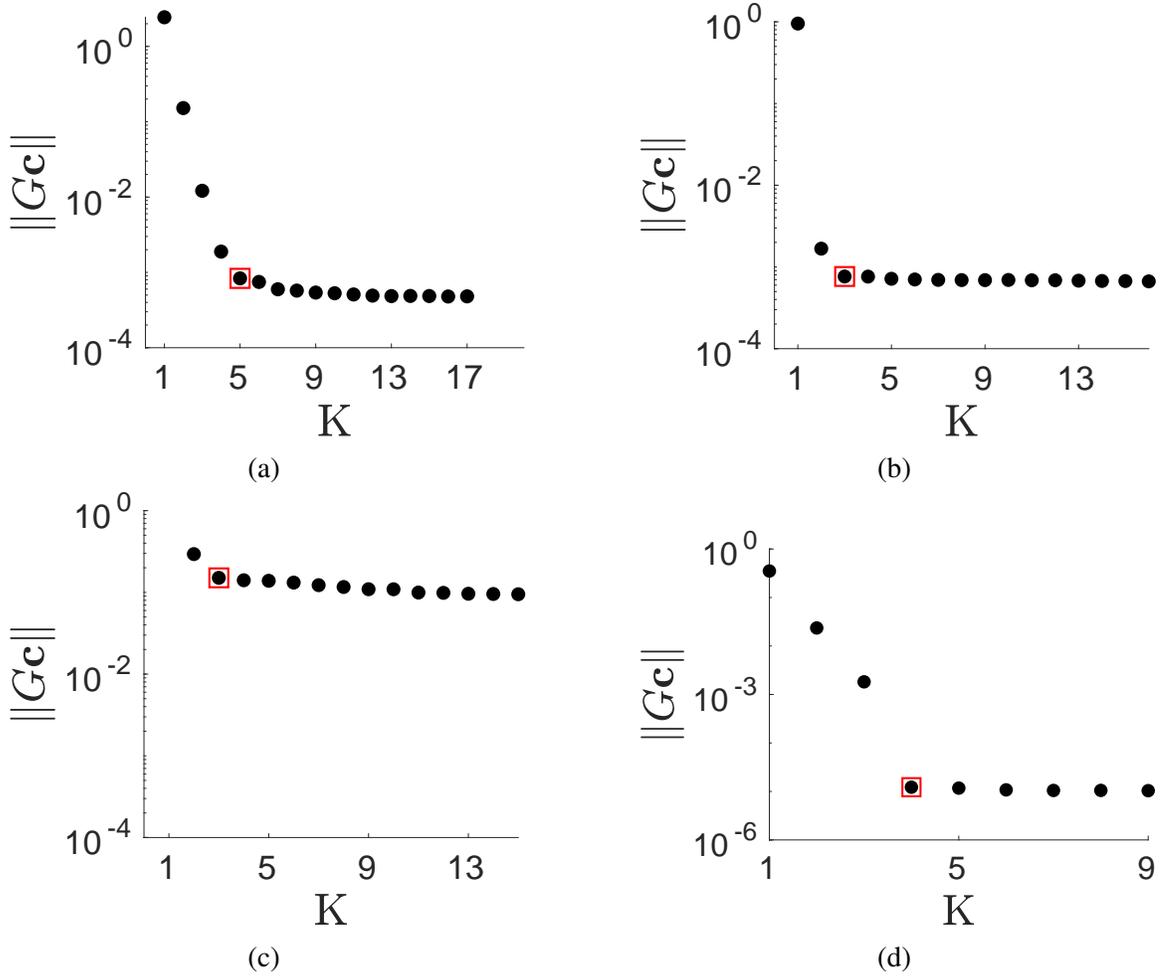


Figure 4.1: **Residual curves for the identified relations.** Panel(a) shows the residual curve for the energy equation. Panel (b) shows the residual curve of the pressure Poisson equation. Panel (c) shows the residual curve of the phenomenological pressure transport relation. Panel (d) shows the residual curve of the Navier-Stokes equation.

of the energy equation: the dissipation term  $(\nabla_i u_j)(\nabla_i u_j)$  must be numerically differenced. We applied a second order centered difference to estimate the derivatives, and this likely caused coefficient errors two orders of magnitude larger than the Navier-Stokes equation. The power of the weak formulation is evident, because these numerical errors still do not prevent the recovery of the governing energy dynamics.

In fact, it is no easy task to recover viscosity in Navier-Stokes when the magnitude of associated contribution to the relation is  $\chi_n = 0.0003$ . This displays the robustness of GISR in discovering equations with terms over a range of magnitudes. Indeed, a small

spurious term proportional to unity is discovered in the pressure Poisson equation with a similar magnitude  $\chi_n = 0.002$  to that of viscous effects in Navier-Stokes. This spurious term is fitting some discretization error arising from integration on non-uniform domains or truncation error in the underlying numerical simulation. It is likely discretization error in the nonuniform direction, since this spurious term is not present if the analysis is repeated using data near the channel wall instead of in the middle of the domain. Furthermore, the identified value of  $c_8$  will fluctuate if a different spacetime region is sampled from near the middle of the channel.

Regardless, the size of this spurious term is very small. The dominant balance  $\nabla^2 p + \nabla_i(u_j \nabla_j u_i) = 0$  can be seen to be a good description of the system in Figure 4.1(b). In fact, this method finds a sequence of dominant balances that describe the system in successively coarser ways. In Figure 4.1(d), one can see the increase in the residual due to discarding terms sequentially from the Navier-Stokes equation. Viscous effects can be neglected to find the Euler equation  $\partial_t u_i + u_j \nabla_j u_i + \nabla_i p = 0$ , and the pressure gradient can be discarded to find Burger’s equation  $\partial_t u_i + u_j \nabla_j u_i = 0$ . The order terms are discarded is determined by their size  $\chi_n$ , with the smallest terms being discarded first. All of these relations are useful physical descriptions of the sampled data identified by SPIDER at different noise levels [64].

## 4.2 Active Nematics

The term “active matter” describes systems made up of self-driven units. Active matter can display unusual emergent phenomena like spontaneous organized motion [65, 66, 67, 68, 69]. In [37], we applied SPIDER to a particular class of such systems, known as active nematics [70]. These systems emerge from collections of many highly elongated apolar interacting units. Examples of active nematics appear in non-biological systems such as vibrated monolayers of cylindrical rods [71]), but our interest is primarily driven by biological systems: actin filaments [72], microtubules [73], and bacteria [74] suspended

in a layer of fluid.

Active nematics exhibit a range of flows featuring topological defects in the orientation field [73, 75, 76, 77] and a number of hydrodynamic models have been proposed in an attempt to understand experimental observations [78, 79, 80, 81, 82, 83, 84, 85, 70, 86, 87]. These models are variations of the locally equivalent Leslie-Ericksen model [88, 89] and explicitly nematic Beris-Edwards model [90, 91], which provide a coarse-grained description of *microscopic* nematic molecules in three spatial dimensions. These models capture some observed phenomena [92], but they also fail to describe a number of experimental observations specifically for MT suspensions [76, 77]. The foremost difficulty for hydrodynamic models of active nematics is that they contain a dozen or so parameters, few of which can be directly measured. This makes it difficult to distinguish the effects of using incorrect parameter values from an altogether incorrect model.

To remedy this situation, SPIDER was applied to two observable fields: the director field  $\mathbf{n}$  and the flow velocity  $\mathbf{u}$ . These observable fields can be seen in Figure 4.2. The presence of topological defects with  $\pm 1/2$  charge assures us this system is nematic invariant:  $\mathbf{n} \equiv -\mathbf{n}$ . In Figure 4.2c, there are discontinuous lines where the director field spontaneously changes sign.  $\mathbf{n}$  can be consistently oriented in any defect-free neighborhood, but there will always be defect-induced discontinuities in any global description. This has interesting consequences for the weak formulation, as will be discussed.

The nematic field  $\mathbf{n}$  is extracted with coherence enhanced diffusion filtering (CEDF) [93, 94]. Figure 4.2b shows that the extracted nematic field is unreliable near defects, so a mask  $\psi$  is constructed to remove defects and their neighborhoods from analysis. There are discontinuous derivatives and systematic noise in our extracted fields at these locations. A snapshot of this mask is shown in Figure 4.2c. The numerical flow field  $\mathbf{u}$  was extracted with Particle Image Velocimetry (PIV) using DaVis software. The flow field and its vorticity can be seen in figure Figure 4.2d. Lengths, times, and velocities have been nondimensionalized such that  $\langle |\mathbf{u}| \rangle = \langle |\nabla \times \mathbf{u}| \rangle = 1$ . Nematic theories are usually accompanied

Table 4.2: **The summary of the model libraries and their symmetry properties.** The nematic symmetry doubles the number of representations worth studying, although two are neglected in this work.

Libraries	nematic invariant	nematic covariant
rank-0 tensor (scalar)	$\mathcal{L}_0$	$\hat{\mathcal{L}}_0$
rank-1 tensor (vector)	$\mathcal{L}_1$	$\hat{\mathcal{L}}_1$
symmetric traceless rank-2 tensor	$\mathcal{L}_2^{STF}$	not studied
antisymmetric rank-2 tensor	$\mathcal{L}_2^{AS}$	not studied

by a scalar order parameter. The microtubule density  $\phi$  is the correct scalar observable, but observations indicate it is saturated away from defects:  $\phi \approx 1$ . Since we are masking low density regions, we exclude  $\phi$  from our analysis.

The symmetry group of the system we will use for library construction is  $O(2) \times \mathbb{Z}_2$ , where  $O(2)$  is the symmetry group of spatial transformations and  $\mathbb{Z}_2$  is the nematic transformation  $\mathbf{n} \rightarrow -\mathbf{n}$ . Since the symmetry group is a direct product, the irreducible representations are tensor products of irreducible representations of the factors. The representations of interest for  $O(2)$  have been investigated earlier: the scalars, vectors, antisymmetric tensors, and symmetric trace-free tensors. The representations of  $\mathbb{Z}_2$  are the nematic invariant (even powers of  $\mathbf{n}$ ) and nematic covariant (odd powers of  $\mathbf{n}$ ). The representations studied are found in Table 4.2.

The first step of library construction is to find all tree tensors, up to a certain rank, from the observables and differential operators. The number of tree tensors grows quickly with rank, so a physics informed cutoff is warranted. The flow is known to be very slow, so inertial effects are certainly negligible. With this in mind, we constrain tree tensors to contain a single power of  $\mathbf{u}$  and  $\partial_t$ . Viscosity and elasticity will manifest as terms with two spatial derivatives, so we also constrain tree tensors to only contain up to two gradients. We will not be interested in terms that mix  $\partial_t$  and  $\nabla$  in this analysis.  $\mathbf{n}$  is an  $O(1)$  field, so

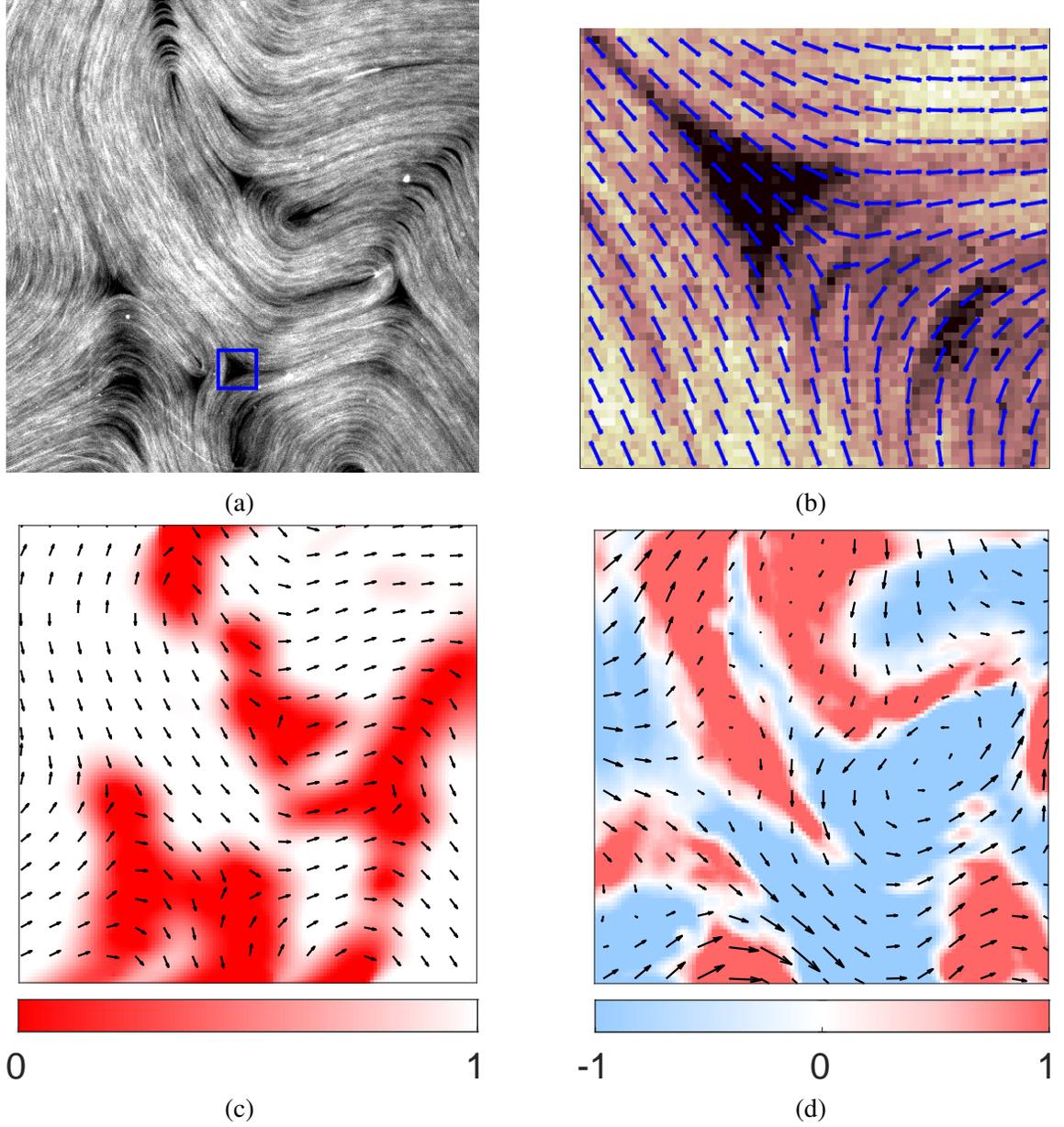


Figure 4.2: **Experimental snapshot and observable fields.** (a) An experimental image of the microtubules. The complete image is shown, with the blue box highlighting a  $-1/2$  topological defect. Panel (b) shows a zoomed-in  $-1/2$  defect and the extracted nematic field around it. The extracted director field  $\mathbf{n}$  (blue arrows) is clearly unreliable near topological defects. (c) Director field  $\mathbf{n}$  (black arrows) and the mask  $\psi$  (color) used to remove near-defect data. Panel (d) shows the flow field  $\mathbf{u}$  (black arrows) and the corresponding vorticity  $\omega = (\nabla \times \mathbf{u})_z$  (color).

no limits are placed on its occurrence in tree tensors. The tree tensors  $\mathcal{T}_{(k)}$  of rank  $k$  are

$$\mathcal{T}_{(0)} \in \{1\},$$

$$\mathcal{T}_{(1)} \in \{\mathbf{u}, \partial_t \mathbf{n}, \partial_t \mathbf{u}\} \cup \{\mathbf{n} \mathcal{T}_{(0)}\}$$

$$\mathcal{T}_{(2)} \in \{\mathbf{u} \partial_t \mathbf{n}, \nabla \mathbf{n}, \nabla \mathbf{u}\} \cup \{\mathbf{n} \mathcal{T}_{(1)}\}$$

$$\mathcal{T}_{(3)} \in \{\nabla \cdot \nabla \mathbf{u}, \nabla \nabla \cdot \mathbf{u}\} \cup \{\mathcal{T}_{(2)}\}$$

If one were to proceed with these tree tensors, a large number of identities would make it into the final libraries  $\mathcal{L}$ . These identities stem from the constraint  $\mathbf{n}^2 = 1$ , which eliminates a degree of freedom from the nematic gradients:

$$n_i \nabla_j n_i = 0, \quad n_i \partial_t n_i = 0. \quad (4.5)$$

This constraint allows the gradient tensor  $\nabla \mathbf{n}$  to be reduced [95], such that the gradient information is described by a splay scalar  $s$  and bend vector  $b_i$

$$\nabla_i n_j = -n_i b_j + s(\delta_{ij} - n_i n_j). \quad (4.6)$$

The splay and bend are uniquely defined by  $s = \nabla \cdot \mathbf{n}$  and  $\mathbf{b} = -(\mathbf{n} \cdot \nabla) \mathbf{n}$ . The splay  $s$  is nematic-covariant,  $b_i$  is nematic-invariant, and  $n_i b_i = 0$ . The tree tensors can be significantly simplified with these objects. We define the concept of reduced tree tensors  $\mathcal{R}_{(k)}$ , which replace gradients  $\nabla_i n_j$  with their lower-rank projections in the case of identities. This has the benefit of reducing the rank of tree tensors including this gradient by at least 1.

$$\begin{aligned} \mathcal{R}_{(0)} &\in \{1, s, s^2\} \\ \mathcal{R}_{(1)} &\in \{\mathbf{u}, \partial_t \mathbf{n}, \partial_t \mathbf{u}, \mathbf{b}, s\mathbf{u}, \nabla s, s\mathbf{b}, s^2 \mathbf{u}\} \cup \{\mathbf{n} \mathcal{R}_{(0)}\} \\ \mathcal{R}_{(2)} &\in \{\mathbf{u} \partial_t \mathbf{n}, \nabla \mathbf{u}, \mathbf{u} \mathbf{b}, \nabla \mathbf{b}, \mathbf{b} \mathbf{b}, s \mathbf{u} \mathbf{b}, s \nabla \mathbf{u}, \mathbf{u} \nabla s\} \cup \{\mathbf{n} \mathcal{R}_{(1)}\} \\ \mathcal{R}_{(3)} &\in \{\nabla \nabla \mathbf{u}, \mathbf{u} \mathbf{b} \mathbf{b}, \mathbf{b} \nabla \mathbf{u}, \mathbf{u} \nabla \mathbf{b}\} \cup \{\mathbf{n} \mathcal{R}_{(2)}\} \\ \mathcal{R}_{(k)} &\in \{\mathbf{n} \mathcal{R}_{(k-1)}\} \text{ for } k > 3. \end{aligned} \quad (4.7)$$

This reduction will not remove all identities from the library, but it eliminates many of them and simplifies the computation of contractions. For evaluation of library terms, it is

convenient to introduce the strain-rate  $A_{ij}$ , vorticity  $\Omega_{ij}$ , and nematic tensor  $Q_{ij}$ .

$$A_{ij} = \frac{1}{2}(\nabla_i u_j + \nabla_j u_i) \quad \Omega_{ij} = \frac{1}{2}(\nabla_i u_j - \nabla_j u_i) \quad Q_{ij} = n_i n_j \quad (4.8)$$

Antisymmetrization as done in  $\Omega_{ij}$  will be denoted by  $T_{[ij]}$ , and the symmetric trace-free part of a tensor can be denoted with a bar  $\bar{A}_{ij} = A_{ij} - (1/2)\delta_{ij}A_{kk}$ .

### Weak-formulation and sparse regression

Integration domains  $\Omega_i$  are chosen with nondimensionalized side length of 2 in space and time. The hydrodynamic length scales are  $\sim 27$  points in space and  $\sim 32$  points in time, so each integration domain consists of  $54 \times 54 \times 65 = 189,540$  data points, where the entire data set consists of  $128 \times 128 \times 530 = 8,683,520$  data points. The number of integration domains is taken to be 10 times the size of the library to ensure  $G$  is sufficiently overdetermined. Four base weight functions  $w_k$  were used in each domain  $\Omega_i$ .

$$\begin{aligned} w_1 &= \psi\chi \prod (1 - \hat{x}_i^2)^4 \\ w_2 &= \psi\chi \cos(\pi x) \prod (1 - \hat{x}_i^2)^4 \\ w_3 &= \psi\chi \cos(\pi y) \prod (1 - \hat{x}_i^2)^4 \\ w_4 &= \psi\chi \cos(\pi t) \prod (1 - \hat{x}_i^2)^4 \end{aligned} \quad (4.9)$$

where  $x$ ,  $y$ , and  $t$  are nondimensionalized and  $\hat{x}_i$  are rescaled to the canonical interval  $[-1, 1]$ . The sinusoidal weights are included to increase sensitivity to variation in the fields.  $\psi$  is the mask shown in Figure 4.2c that masks defect neighborhoods, and  $\chi$  is a topological mask that is independently treated for each library. For nematic invariant libraries  $\chi = 1$ , and for nematic covariant libraries,  $\chi$  will be an odd function of  $\mathbf{n}$ .

Model discovery for 3D turbulence data was performed by greedy minimization of  $\|G\mathbf{c}\|_2$  in sparse subspaces via GISR. For this system, we will instead use the non-convex

relative residual Equation 3.11 to rank model performance. Physical relations ought to have small values for both the absolute residual and the relative residual. To achieve both goals, the SVD is used to estimate coefficients to minimize the absolute residual, but model modifications are selected by their impact on the relative residual. Since this residual definition is not useful for single term models, a modification is made to the denominator when a single term remains.

$$\eta_{\text{single-term}}(c_n) = \frac{\|\mathbf{g}_n\|_2}{\|\mathbf{H}\|_2} \quad (4.10)$$

$\mathbf{g}_n$  is the column of  $G$  associated with the single term, and the denominator has been replaced with a positive definite scale  $\mathbf{H}$ , which is the weak form the Frobenius norm of the tree tensor used in constructing  $\mathbf{F}_n$ . For example, the normalization for incompressibility  $\nabla_i u_i = 0$  will be the weak evaluation

$$H_k = \frac{1}{N_k S_H} \int_{\Omega_k} \sqrt{(\nabla_i u_j)(\nabla_i u_j)}, \quad (4.11)$$

where the normalization factors  $N_k$  and  $S_H$  are identical to those of Equation 3.1. The off diagonal components of  $\nabla_i u_j$  are included in this scale estimation since the sum with diagonal components alone  $(\partial_x u_x)^2 + (\partial_y u_y)^2 + (\partial_z u_z)^2$  is not a rotational scalar. If proper nondimensionalization has been done, these single-term scales should be  $O(1)$ . Regression is halted when  $\eta^{(n-1)} > \gamma \eta^n$  with  $\gamma = 1.15$ .

#### Nematic-invariant scalar library

There are nine nematic-invariant scalars that can be obtained from even-rank reduced fundamental tensors with the same nematic symmetry listed in (Equation 4.7):

$$\mathcal{L}_0 = \{1, s^2, n_i u_i s, \nabla_i(n_i s), \nabla_i u_i, \bar{Q}_{ij} \bar{A}_{ij}, u_i b_i, b_i b_i, \nabla_i b_i\} \quad (4.12)$$

Nematic invariant libraries can be integrated with  $\chi = 1$ . This library contains a single identity that is identified by GISR with low residual.

$$\nabla_i(n_i s + b_i) = 0 \quad (4.13)$$

The latter term  $\nabla_i b_i$  is then pruned. Two parsimonious physical relations are identified via sparse regression, the incompressibility condition (Equation 4.14) and a relation (Equation 4.15) between the director and flow fields

$$\nabla_i u_i = 0 \quad (4.14)$$

$$\bar{Q}_{ij} \bar{A}_{ij} + c_1^{(1)} = 0 \quad (4.15)$$

where  $c_1^{(1)} = -0.55 \pm 0.3\%$  with the relative residuals of  $\eta = 0.03$  and  $\eta = 0.08$ , respectively. Note that the coefficient  $c_1$  has units of inverse time, and it defines a timescale comparable to the natural hydrodynamic timescale used in nondimensionalization. Since  $\nabla_i u_i$  is found to be zero in the region we are sampling, we prune it from all subsequent libraries. One could in principal also prune  $\bar{Q}_{ij} \bar{A}_{ij}$ , but since this relation is nonstandard, we keep it to see if a generalization or contradiction is found.

#### Nematic-covariant scalar library

There are 18 nematic-covariant scalars that can be constructed from even-rank reduced fundamental tensors:

$$\begin{aligned} \hat{\mathcal{L}}_0 = \{ & s, n_i u_i, n_i \partial_t u_i, u_i \partial_t n_i, n_i u_i s^2, s u_i b_i, s \nabla_i u_i, s \bar{Q}_{ij} \bar{A}_{ij}, u_i n_i n_j \nabla_j s, u_i \nabla_i s, \\ & n_i n_j n_k \nabla_i \nabla_j u_k, n_i \nabla^2 u_i, n_i \nabla_i \nabla_j u_j, u_i n_i b^2, b_i n_j \bar{A}_{ij}, b_j n_i \Omega_{ij}, u_i n_j \nabla_j b_i, n_i u_i \nabla_j b_j \}, \end{aligned} \quad (4.16)$$

where the identity (Equation 4.13) can be used to prune the last term from the library. All terms in the nematic-covariant scalar library  $\hat{\mathcal{L}}_0$  have the same branch cuts as  $\mathbf{n}$ . Figure 4.3(a) shows the inherited branch cuts in the splay scalar in the lower half of the image. Model discovery with discontinuous data requires special care, as discussed in section 3.7. Luckily, these discontinuities can be removed exactly using the topological weight  $\chi$  mentioned in Equation 4.9. We take  $\chi = \mathbf{n}$ ; each term is multiplied by  $\mathbf{n}$  to create nematic invariant vectors. Figure 4.3(b) shows the x-component of this weight function smoothing a line discontinuity. This definition of  $\chi$  effectively embeds  $\hat{\mathcal{L}}_0$  as a subspace of the nematic

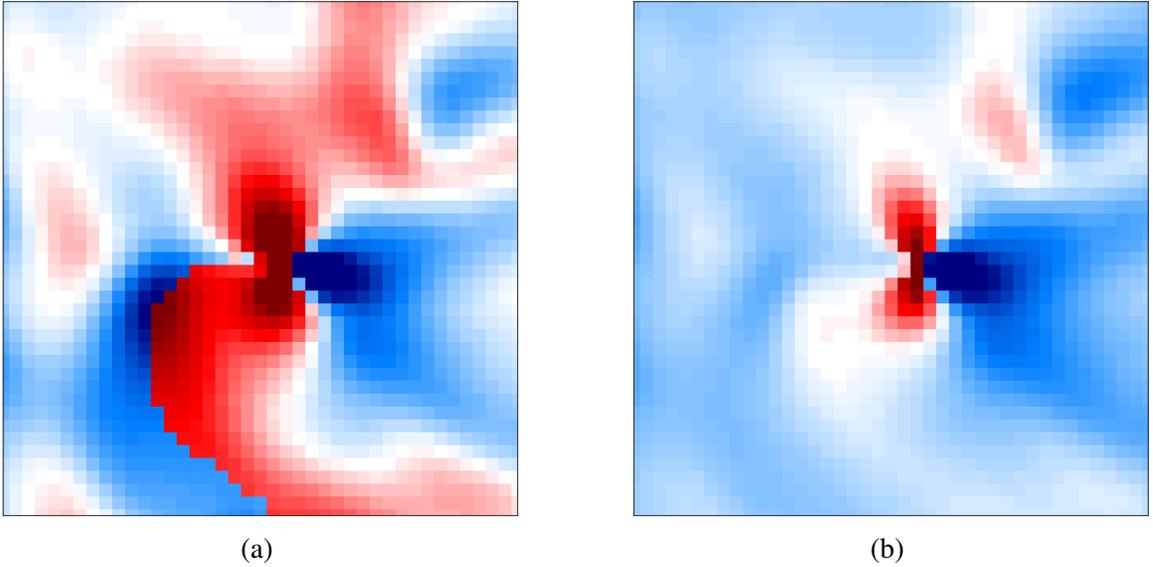


Figure 4.3: **Remedying the branch cuts of the nematic field for weak-form evaluation.** Panel (a) shows the splay scalar  $s$  around a topological defect and a visible branch cut. Panel (b) shows  $sn_x$ , which is nematic invariant and no longer suffers from a line discontinuity. The colorbar is arbitrary.

invariant  $\mathcal{L}_1$ . Only one relation is found in this library:

$$s \left[ \bar{Q}_{ij} \bar{A}_{ij} + c_1^{(2)} \right] = 0, \quad (4.17)$$

where  $c_1^{(2)} = -0.57 \pm 1\%$ . This relation is equivalent to Equation 4.15 as long as  $c_1^{(2)} = c_1^{(1)}$ ,

### Nematic-covariant vector library

The nematic-covariant vector library should contain the nematodynamics relation including  $\partial_t n_i$ . Since  $|\mathbf{n}| = 1$ , this time derivative should be orthogonal to  $\mathbf{n}$ , so without loss of generality, we can restrict our library entirely to vectors orthogonal to  $\mathbf{n}$ .

$$\hat{\mathcal{L}}_1 = \hat{P}_\perp \{ \partial_t n_i, s u_i, \nabla_i s, s b_i, \bar{A}_{ij} n_j, \Omega_{ij} n_j, u_j \nabla_j n_i, n_j \nabla_j b_i \}, \quad (4.18)$$

where the term  $n_j u_j b_i$  has been replaced by its more familiar form  $u_j \nabla_j n_i$  appearing as the advective nonlinearity of the Ericksen-Leslie equation.  $\hat{P}_\perp$  projects vectors to be orthogonal to  $n_i$ . The topological mask  $\chi$  is taken to be the *cross product with the nematic field*  $\chi(\mathbf{f}) = \mathbf{f} \times \mathbf{n}$ , since this simultaneously resolves the branch cuts and projects the library onto the subspace orthogonal to  $\mathbf{n}$ . No identities are found in this library. Sparse regression identifies a single parsimonious physical relation (Equation 4.19).

$$\partial_t n_i + c_2^{(1)} u_j \nabla_j n_i + c_3^{(1)} \Omega_{ij} n_j + c_4^{(1)} \hat{P}_\perp \bar{A}_{ij} n_j = 0, \quad (4.19)$$

where  $c_2^{(1)} = (0.99 \pm 0.8\%)$ ,  $c_3^{(1)} = (-0.95 \pm 0.7\%)$ , and  $c_4^{(1)} = (-0.95 \pm 1\%)$ . This relation is formally equivalent to the evolution equation of the Leslie-Ericksen model [89, 88, 96] without relaxation of elastic energy via rotational diffusion with coefficients  $c_r$  that are very close to  $\pm 1$ . The relative residual  $\eta = 0.08$  is quite low and comparable to that of equation Equation 4.15.

### Nematic-invariant vector library

The nematic-invariant vector library would be expected to include a momentum balance relation, which contains divergences of various stresses. In order to allow elastic effects,

this library is allowed to contain divergences of rank-2 tensors  $\nabla_j \sigma_{ij}$ .

$$\begin{aligned}
\mathcal{L}_1 = & \{sn_i, u_i, (n_j u_j)n_i, \partial_t u_i, n_i n_j \partial_t u_j, b_i, u_i s^2, n_i n_j u_j s^2, n_j u_j \partial_t n_i, n_i u_j \partial_t n_j, sn_j u_j b_i, \\
& su_j b_j n_i, n_i s \bar{Q}_{jk} \bar{A}_{jk}, sn_j \nabla_i u_j, sn_j \nabla_j u_i, sn_i \nabla_j u_j, n_i (n_j u_j) (n_k \nabla_k s), u_i n_j \nabla_j s, \\
& n_j u_j \nabla_i s, n_j n_k \nabla_j \nabla_k u_i, n_j n_k \nabla_i \nabla_j u_k, u_i b^2, b^2 u_j n_j n_i, u_j b_j b_i, b_i \bar{A}_{jk} \bar{Q}_{jk}, b_i \nabla_j u_j, \\
& b_j \nabla_j u_i, b_j \nabla_i u_j, n_j u_j n_k \nabla_k b_i, u_i \nabla_j b_j, u_j \nabla_i b_j, u_j \nabla_j b_i\} \\
& \cup \nabla_i \{ \bar{Q}_{ij}, s^2 \bar{Q}_{ij}, su_i n_j, su_j n_i, su_k n_k \bar{Q}_{ij}, (n_k \nabla_k s) \bar{Q}_{ij}, n_i \nabla_j s, n_j \nabla_i s, n_i s b_j, n_j s b_i, \\
& (\nabla_k u_k) \bar{Q}_{ij}, (\bar{A}_{kl} \bar{Q}_{kl}) \bar{Q}_{ij}, (\nabla_i u_k) n_k n_j, (\nabla_j u_k) n_k n_i, \bar{A}_{ij}, \Omega_{ij}, (u_k b_k) \bar{Q}_{ij}, u_k n_k b_i n_j, \\
& u_k n_k b_j n_i, u_i b_j, u_j b_i, (\nabla_k b_k) \bar{Q}_{ij}, (n_k \nabla_k b_i) n_j, (n_k \nabla_k b_j) n_i, \nabla_i b_j, \nabla_j b_i, b^2 \bar{Q}_{ij}, b_i b_j \}
\end{aligned} \tag{4.20}$$

This is the largest library considered compared to the other irreducible representations. Although we are including all possible stress tensors, the contribution from pressure  $p\delta_{ij}$  is a latent field and was not experimentally observed. The effect of this latent field can be removed by integrating the curl of the vector library [35, 36]. This requires one to numerically differentiate the mask  $\psi$  at least once for every integration. This highlights the importance of a smooth mask function.

There are a number of identities in this library that will not be listed explicitly. Symbolic regression identified two parsimonious relations:

$$(\bar{A}_{kl} \bar{Q}_{kl} + c_1^{(3)}) sn_i + \nabla_i p = 0, \tag{4.21}$$

$$\nabla_k \left[ (\bar{A}_{lm} \bar{Q}_{lm} + c_1^{(4)}) \bar{Q}_{ik} + \delta_{ik} p \right] = 0, \tag{4.22}$$

where  $c_1^{(3)} = -0.57 \pm 1\%$  and  $c_1^{(4)} = -0.59 \pm 1\%$ . The relative residual of these relations  $\eta = 0.28$  and  $\eta = 0.38$ , respectively, which is due to the high number of derivatives needed to evaluate these relations.

### Symmetric trace-free tensor library

It is not always convenient to use a bar to denote the symmetric trace-free part of a tensor. For this section, let us override the symmetrization notation such that  $T_{(ij)}$  denotes the trace-free symmetric part  $T_{(ij)} = \frac{1}{2}(T_{ij} + T_{ji} - \delta_{ij}T_{kk})$ .

$$\begin{aligned} \mathcal{L}_2^{STF} = \{ & \bar{Q}_{ij}, s^2\bar{Q}_{ij}, \partial_t\bar{Q}_{ij}, su_{(i}n_{j)}, su_k n_k \bar{Q}_{ij}, (n_k \nabla_k s)\bar{Q}_{ij}, n_{(i}\nabla_{j)}s, sn_{(i}b_{j)}, \\ & \nabla_k u_k \bar{Q}_{ij}, (\bar{A}_{kl}\bar{Q}_{kl})\bar{Q}_{ij}, \bar{A}_{ij}, (u_k b_k)\bar{Q}_{ij}, u_k n_k b_{(i}n_{j)}, (\nabla_k b_k)\bar{Q}_{ij}, \\ & (n_k \nabla_k b_{(i}n_{j)}), (\nabla_i b_j)', b^2\bar{Q}_{(ij)}, \bar{A}_{k(i}\bar{Q}_{j)k}, b_{(i}b_{j)}, u_{(i}b_{j)} \}. \end{aligned} \quad (4.23)$$

Note that symmetric trace-free tensors have two independent components  $(i, j) = (1, 1)$  and  $(1, 2)$ , doubling the number of rows in  $G$ . Three symbolic identities appear in this library. They are found with exceptionally low residuals.

$$\begin{aligned} \bar{A}_{k(i}\bar{Q}_{j)k} &= 0, \\ b_{(i}b_{j)} + b^2\bar{Q}_{(ij)} &= 0, \\ u_k \nabla_k Q_{(ij)} - 2su_{(i}n_{j)} + 2su_k n_k Q_{(ij)} + 2u_k b_k \bar{Q}_{(ij)} + 2u_{(i}b_{j)} &= 0. \end{aligned} \quad (4.24)$$

We use these to discard the last two library terms. Two physical relations are found in this library, the stress balance relation Equation 4.25 and an evolution equation for the orientation tensor Equation 4.26.

$$\bar{A}_{kl}\bar{Q}_{kl}Q_{ij} + c_1^{(5)}Q_{ij} = 0, \quad (4.25)$$

$$\partial_t Q_{ij} + c_2^{(2)}u_k \nabla_k \bar{Q}_{ij} + c_3^{(2)}(\Omega_{ik}\bar{Q}_{kj} - \bar{Q}_{ik}\Omega_{kj}) + c_4^{(2)}\bar{A}_{ij} + c_5^{(2)}Q_{ij}(\bar{A}_{kl}\bar{Q}_{kl}) = 0, \quad (4.26)$$

where  $c_1^{(2)} = 1 \pm 0.1\%$ ,  $c_2^{(2)} = -0.96 \pm 0.1\%$ ,  $c_3^{(2)} = -1.02 \pm 0.1\%$ , and  $c_4^{(2)} = 2.05 \pm 0.1\%$ . These relations have low residuals  $\eta = 0.1$  and  $\eta = 0.09$ , respectively. For comparison, the tensor balance between  $\bar{A}_{ij}$  and  $\bar{Q}_{ij}$  proposed in Ref. [82] has a much higher residual

$\eta = 0.67$ .

### Antisymmetric tensor library

Let  $T_{[ij]} = \frac{1}{2} (T_{ij} - T_{ji})$  be the antisymmetric part of a rank-2 tensor. Again, we can construct the library of antisymmetric tensors using even-rank reduced fundamental tensors:

$$\begin{aligned} \mathcal{L}_2^{AS} = \{ & n_{[i} \partial_t n_{j]}, su_{[i} n_{j]}, n_{[i} \nabla_{j]} s, sn_{[i} b_{j]}, n_k n_{[i} \nabla_{j]} u_k, \Omega_{ij}, u_k n_k b_{[i} n_{j]}, u_{[i} b_{j]}, \\ & n_{[i} n_k \nabla_k b_{j]}, \nabla_{[i} b_{j]} \}. \end{aligned} \quad (4.27)$$

This library can be handled in the same way as the nematic-invariant scalar library. In fact, each tensor has a single degree of freedom (also known as the pseudo-scalar) making the construction of  $G$  completely analogous. We find a single identity

$$u_k n_k b_{[i} n_{j]} + u_{[i} b_{j]} = 0 \quad (4.28)$$

and a single physical relation again corresponding to the nematodynamics.

$$n_{[j} \partial_t n_{i]} + c_2^{(3)} su_{[i} n_{j]} + c_6^{(3)} u_{[i} b_{j]} + c_3^{(3)} \Omega_{ij} + c_4^{(3)} n_k n_{[i} \nabla_{j]} u_k = 0, \quad (4.29)$$

where  $c_2^{(3)} = 1.03 \pm 0.2\%$ ,  $c_6^{(3)} = 1.04 \pm 0.9\%$ ,  $c_3^{(3)} = -0.98 \pm 0.5\%$ , and  $c_4^{(4)} = -1.00 \pm 0.7\%$ . The relative residual for relation (Equation 4.29) is  $\eta = 0.05$ , making it the most accurate representation of  $\partial_t \mathbf{n}$  that was found.

### Discussion

Six tensor libraries were constructed from the two vector fields  $\mathbf{u}$  and  $\mathbf{n}$ . Evaluating library terms and performing sparse regression led to the discovery of nine relations, with at least one in each tensor space investigated. Figure 4.4 shows the sparsity curves of all discovered relations.

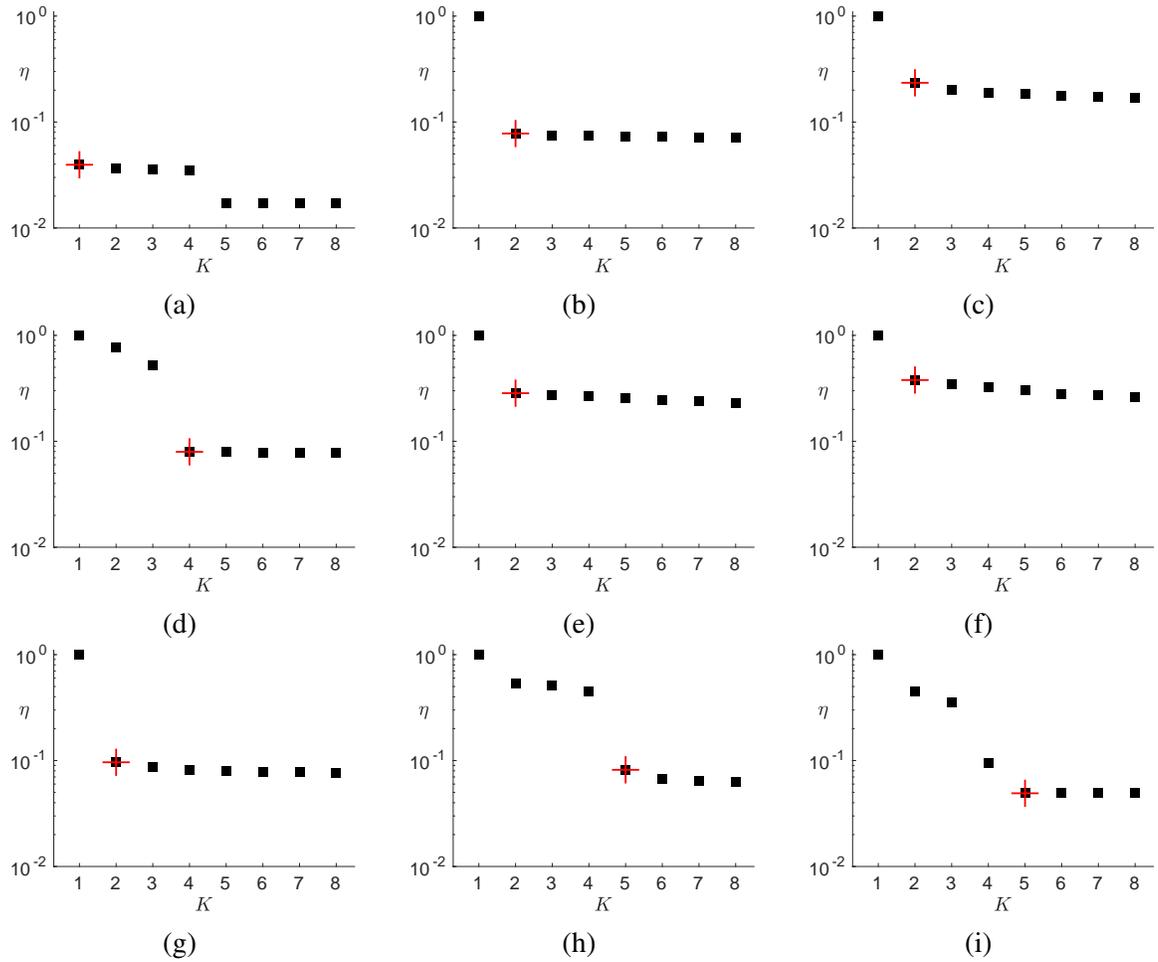


Figure 4.4: **The relative residual curve as a function of the number of terms in the relation.** The plots corresponding to discovered equations are (a) (Equation 4.14), (b) (Equation 4.15), (c) (Equation 4.17), (d) (Equation 4.19), (e) (Equation 4.21), (f) (Equation 4.22), (g) (Equation 4.25), (h) (Equation 4.26), and (i) (Equation 4.29). The red cross indicates the corresponding parsimonious relation.

These nine relations are equivalent to a set of three: the incompressibility condition Equation 4.14, the nematodynamics relation Equation 4.19, and the constant extension relation Equation 4.15. The physical interpretations of these equations is discussed in [37] but will be briefly recapped here. The minimal set of equations are presented below for reference. Note that this is a complete model of the fields  $\mathbf{u}$  and  $\mathbf{n}$  when the density is saturated.

$$\nabla_i u_i = 0$$

$$\bar{A}_{ij} \bar{Q}_{ij} + c_1 = 0$$

$$\partial_t n_i + c_2 u_j \nabla_j n_i + c_3 \Omega_{ij} n_j + c_4 \hat{P}_\perp \bar{A}_{ij} n_j = 0$$

SPIDER only investigates the relations in weak form. Strong form images of these relations can be seen in Figure 4.8. The incompressibility condition represents mass conservation, and it is nontrivial that the 2D divergence effectively vanishes away from defects. Since the incompressibility condition is expected to hold for the full 3D flow, this tells us that  $\partial_z u_z$  is very small away from defects, and since  $u_z \approx 0$  at the interface, the flow is truly effectively 2D in the regions studied. However, the strong form suggests this is not the case near defects, especially  $+1/2$  defects. Visual inspection shows positive divergence at the heads of such defects, which indicates  $\partial_z u_z \neq 0$  there. The flow must become 3D in some way at these locations, and the 2D incompressibility condition must be generalized. The nematodynamics equation has the lowest residual in its antisymmetric rank 2 representations (equivalent to  $\varepsilon_{ij} \partial_t \theta$ ) and appears to capture the evolution of the flow field well. Results of forecasting using this equation can be seen in Figure 4.5. The strong form of nematodynamics Figure 4.8(c-d) holds well everywhere, including near defects. It likely does not need to be generalized near defects, although it should probably be cast in terms of a weighted director field  $\phi \mathbf{n}$  where  $\phi$  is the microtubule density. It is notable that no

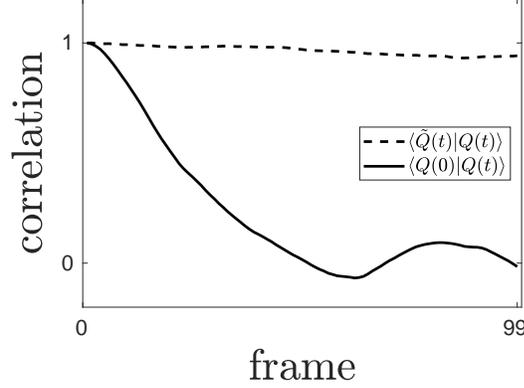


Figure 4.5: **Correlation between the predicted nematic field and the observed nematic field.**  $Q(t)$  is the experimentally observed field and  $\tilde{Q}(t)$  is numerically integrated via the leapfrog method from the first two frames using the state velocity predicted by Equation 4.19 using the experimental flow field.  $\langle f|g \rangle$  is the normalized inner product in the sense of Dirac. The lowest correlation between numerical predictions and experiment is 0.93. This minimum lowers to 0.88 if the constant extension constraint is approximately solved using experimental boundary conditions for the streamfunction  $\psi$ . Note that good correlation is found even after the correlation with the initial state has completely vanished.

rotational viscosity<sup>1</sup> is identified in our analysis in any of the three independently identified forms of nematodynamics.

Equation 4.15 was an unexpected result with a nice physical interpretation.  $\mathcal{E} = -c_1$  has units of  $s^{-1}$ : it is the extension rate of the flow. In the corotating rest frame of a point<sup>2</sup>, the flow is locally extensible with an effectively constant extension rate  $\mathcal{E}$ . The flow contracts perpendicular to  $\mathbf{n}$  and expands parallel to  $\mathbf{n}$ . This quantity is directly observable as opposed to the activity coefficient  $\alpha$  used to define the active stress  $\sigma_{ij}^{active} = \alpha Q_{ij}$ . If one uses incompressibility to define the streamfunction  $\psi$  via  $\vec{u} = \partial_y \psi \hat{i} - \partial_x \psi \hat{j}$ , then Equation 4.15 becomes a second order constraint on  $\psi$ .

$$[2\bar{Q}_{xx}\partial_x\partial_y + \bar{Q}_{xy}(\partial_y^2 - \partial_x^2)]\psi = \tilde{g}^{ij}\partial_i\partial_j\psi = \mathcal{E} \quad (4.30)$$

$\tilde{g}^{ij}$  is an effective metric of signature  $(+1, -1)$ , so this equation is a wave equation with a

<sup>1</sup>Rotational viscosity is direct coupling of  $\partial_t \mathbf{n}$  to second spatial derivatives of the director like  $\nabla^2 \mathbf{n}$ .

<sup>2</sup>The comoving, corotating frame is defined by  $x_i(t) = x_i - u_i t - \Omega_{ij} x_j$ , where  $u_i$  and  $\Omega_{ij}$  are the instantaneous velocity and vorticity.

constant source term. This is strikingly different from the usual Stokes

$$\mu \nabla^2 u_i + \nabla_j (\alpha Q_{ij}) = 0 \quad (4.31)$$

used in many effective models of the flow [81]. In the Stokes relation, the differential operator acting on the flow is isotropic and the source term that introduces anisotropy via the divergence of the nematic tensor. The Stokes relation when applied as an effective 2D model makes many assumptions: for example the vanishing of any anisotropic viscous stress, which is described via Leslie viscosities [97]. Suppose  $\mathbf{n}$  points along the  $x$ -axis such that  $Q_{xx} = 1/2$  and  $Q_{xy} = 0$ . The solution to Equation 4.30 is then

$$\psi(x, y) = \mathcal{E}xy + f(x) + g(y), \quad (4.32)$$

where  $f$  and  $g$  capture boundary conditions. An immediate difficulty in solving this relation numerically for nontrivial  $\mathbf{n}$  is that not all boundary conditions for  $\psi$  are compatible with solutions. Even for the trivial case of uniform  $\mathbf{n}$ ,  $f(x)$  and  $g(y)$  are completely determined by boundary conditions on two of the four walls of any rectangular domain.

Numerical solutions to Equation 4.30 with finite difference schemes lead to high frequency oscillations of  $\psi$ . Overdetermining Equation 4.30 with a simultaneous regularization condition  $\nabla^2 \psi = 0$  effectively minimizes the integral of  $|\nabla \psi|^2 = |\mathbf{u}|^2$ . This regularization is weighted by a small parameter  $\lambda$  to ensure the constant extension constraint take numerical priority. Solving these in the least squares sense produces smooth stream-functions that match the constant extension requirement as well as possible. Figure 4.7 shows the results of numerically solving the regularized constraint with  $\lambda = 0.001$  on a circle of radius 1 with no-slip boundary conditions. *Trivially*, the flow fields are consistent with the symmetries of the driving nematic fields, although some symmetry breaking by the numerical grid is present in Figure 4.7(b). *Nontrivially*, the predicted vortex structure qualitatively matches the vortices observed in experiment for the  $+1/2$  defect. This can be

compared to flow fields derived by Giomi et al. with the Stokes law Equation 4.31 [79, 81]. The predicted flow for the  $-1/2$  defect is quite different than Giomi's, and flow pattern nearest to the defect is not observed in experiment. Note that since there is no length scale in our identified equations, the length scale of these simulated flows is set by the radius of the spatial domain. However, in multi-defect configurations the average defect separation would induce a length scale to facilitate comparison with experiment.

Another insight into these relations can be seen in Figure 4.6. The velocity gradient  $\nabla_i u_j$  can be uniquely decomposed into four SO(2) scalars in the presence of a director field: the divergence  $\nabla \cdot \mathbf{u}$ , the vorticity  $\nabla \times \mathbf{u}$ , the nematic aligned extension rate  $\mathbf{n} \cdot \bar{A} \cdot \mathbf{n}$ , and the nematic skewed extension rate  $\mathbf{n} \times \bar{A} \cdot \mathbf{n}$ . The histogram shows that while the vorticity has an enormously wide distribution, the divergence and nematic aligned extension rate are narrow Gaussian-like distributions. The divergence is centered at zero, while the nematic aligned extension rate is centered around 0.55, as predicted by Equation 4.15. This nematic aligned extension rate is the only distribution that displays explicit time-reversal symmetry breaking, since time reversal changes the sign of the velocity gradients. The fact that these two distributions have similar shapes implies two things. (i) They are not perfect descriptions and need to be generalized near defects. (ii) The similarity in the width of the distributions implies they are valid in the same amount of data, which is consistent with their validity in the same dense nematic regions.

This analysis assumed microtubule density  $\phi = 1$  is constant in regions considered. Future work must incorporate  $\phi$  into the model. The dynamics (or constraint) of  $\phi$  should come from the generalization of the incompressibility condition. Since  $\phi$  is certainly locally conserved, generalized models should take the form of continuity equations  $\partial_t \phi + \nabla_i j_i = 0$ , where  $j_i$  is some density current. The constant extension rate will also be generalized, perhaps to a non-scalar equation to allow simultaneous determination of  $\phi, \mathbf{u}, \mathbf{n}$ . Work by Daiyue Sun and myself attempted to extract the density by numerically solving the system of equations  $\partial_t \phi + \nabla_i (u_i \phi) = 0$  or  $\partial_t \phi + u_i \nabla_i \phi$  with GMRES. The field  $\phi$  was

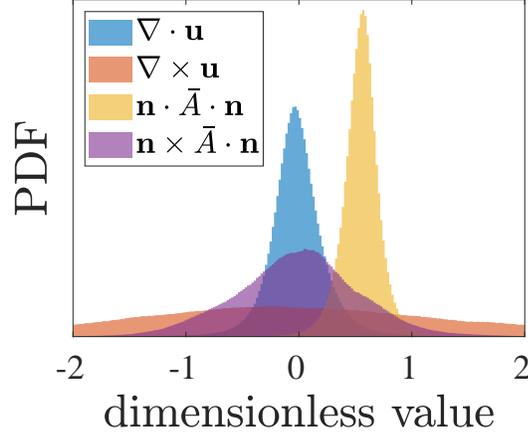


Figure 4.6: **Histogram of flow gradient scalars.** The boundaries in space and time have been removed, as well as regions with  $\psi < 0.2$ . The scale of the PDF is arbitrary but consistent.

subject to the over-determined conditions of mass conservation and proportionality to the experimental image intensity. These efforts were unsuccessful; the resultant field had either unphysical variation or a high residual for the continuity equation. It is an open problem to experimentally measure the microtubule density, even in a normalized sense. Beyond this difficulty, there are also density fields for the ATP and kinesin motor proteins that have not been measured. In weak active turbulence with low defect density as studied here, we believe the assumption of uniform ATP and kinesin density is a good approximation.

An independent model discovery analysis of active nematic turbulence was carried out by Joshi et al. on an independent experimental setup [98]. Their study has a number of key differences with my work.

1. While their library was physically motivated, rotational symmetry was not enforced. For example, coefficients for  $u_x \partial_x \mathbf{Q}$  and  $u_y \partial_y \mathbf{Q}$  were not forced to be equal.
2. They examined multiple data sets with varying activity, while our study only focused on a single timeseries.
3. Their models were ranked with the statistical error  $1 - R^2$ , in contrast to our relative residual  $\eta$ .

4. Their library for the vector equation contained seven terms, in contrast to our  $O(60)$ . The library was constructed ad hoc and excluded all elastic effects.
5. Their analysis assumed that  $\nabla^2 \mathbf{u}$  played an important role in the flow dynamics. This makes their method incapable of discovering our constant extension constraint. This highlights the importance of homogeneous regression in model discovery.

Joshi et al. found the same nematodynamics relation Equation 4.19, which is likewise lacking rotational viscosity. Their regression settled on the curl of Equation 4.31 as the best momentum relation. While is not equivalent to our constant extension relation, their results do corroborate the lack of a length scale.

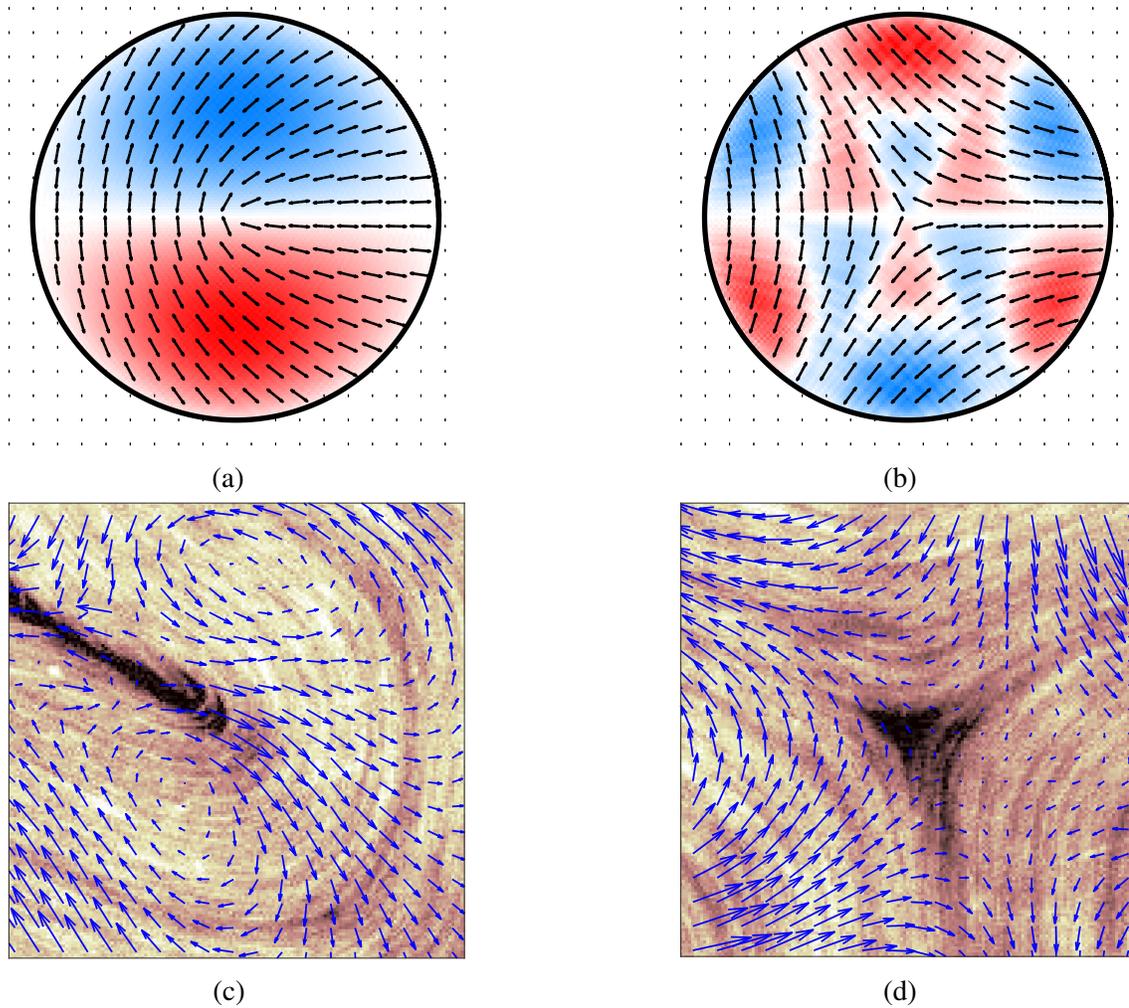


Figure 4.7: **Predicted flow around confined topological defects.** (a) An ideal  $+1/2$  defect generates two counter-rotating vortices responsible for self-driven motion. (b) An ideal  $-1/2$  defect generates a hexagonal flow. (c) An experimental flow field in the comoving, corotating frame of a  $+1/2$  defect. Two counter-rotating vortices are visible. (d) An experimental flow field in the comoving, corotating frame of a  $-1/2$  defect. Alternating vortices are visible, although perfect hexagonal symmetry is not realized. The velocity is shown in the comoving, corotating frame since in such a frame  $A_{ij}$  locally dominates the flow structure.

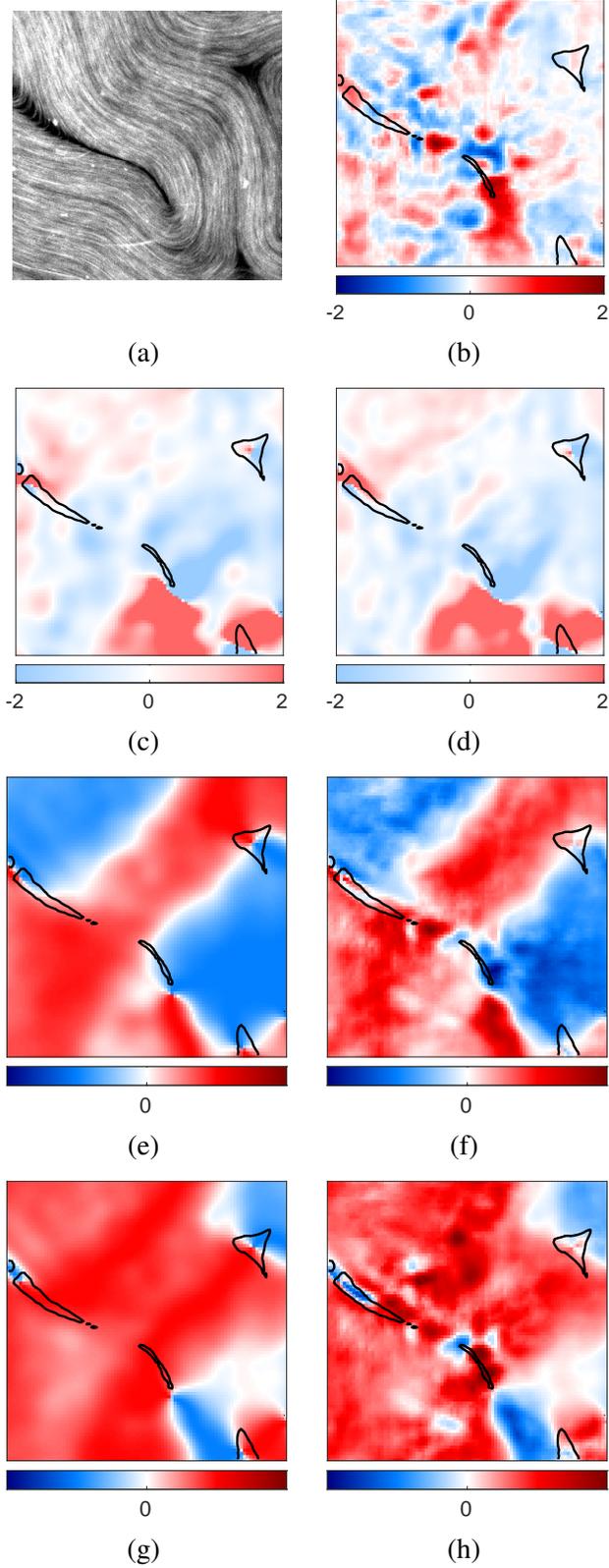


Figure 4.8: **Verification of relations in strong form.** Panel (a) shows the experimental frame of interest. Panel (b) shows the strong form of  $\nabla_i u_i$ , which has fluctuations over small length scales but no recognizable structure away from  $+1/2$  defects. Panel (c) shows the time derivative  $\partial_t \theta = \mathbf{n} \times \partial_t \mathbf{n}$  computed by finite differencing the director field in time, and Panel (d) shows that same quantity, but calculated with spatial finite differencing from Equation 4.19. Panels (e) and (f) show  $c_1^{(1)} Q_{xx}$  and  $\bar{A}_{kl} \bar{Q}_{kl} Q_{xx}$  of Equation 4.25. Panels (g) and (h) show the corresponding  $xy$  components.

## CHAPTER 5

### CONCLUSION

Sparse regression is a powerful tool for equation discovery. SPIDER is a realization of symbolic regression specifically designed for discovering physical laws and effective approximations of spatially extended systems. SPIDER is a *systematic procedure*, with the physicist specifying symmetries of the problem and a physically-informed cutoff for library generation. Physics domain knowledge enters into SPIDER in three ways. Assumptions of locality, smoothness, and translational symmetry constrain PDEs to be local sums tensors with constant coefficients. Physics knowledge will inform library cutoffs (i.e. assuming linearity in small valued fields). Physics-informed nondimensionalization improves regression and allows terms across many scales to be identified. Knowledge of symmetry also plays a critical role. Enforcing symmetry covariance shrinks libraries and increases the number of available measurements for identifying each relation. Irreducible representations prevent the unknowing addition of bad physics into an otherwise accurate relations: each idea learned about a system is learned independently to reduce bias. The weak formulation of differential equations increases robustness to noise and sidesteps numerical differentiation in many scenarios. Discontinuities can be masked from analysis, and the effect of some unobserved fields can be completely removed. Lastly, homogeneous regression places all library terms on equal footing so that no unwarranted assumptions are made about the form of models.

This work has explicitly demonstrated the application of SPIDER to two systems with spatiotemporal variation: 3D fluid turbulence and 2D active nematic turbulence. SPIDER recovered a complete description of the system in both cases and provided physical insight through effective relations and dominant balances. In the case of active nematics, the constant extension relation has not been derived from first principles, but it was found

independently four times as the best relation in various tensor libraries. Future work should incorporate the microtubule density into models.

SPIDER seems particularly well-poised to discover emergent hydrodynamic models of discrete interacting units if simulations or experiments are feasible. Work is ongoing to find PDE descriptions of flocking produced by the microscopic Vicsek model [99]. The results of SPIDER can be compared with existing hydrodynamic predictions [100, 101, 102]. Even more fundamental systems like collections of particles interacting via Newton’s laws should be investigated. SPIDER should be able to learn the equation of state through momentum transport relations from a sufficient number of particles. Perhaps symmetry-covariant corrections to Navier-Stokes can be learned in the low particle, high Knudsen number limit. Promising work in this direction has been done on kinetic simulations of plasmas [103] and interacting particles with Brownian noise [27].

While SPIDER was built for physics, symbolic regression on covariant tensor libraries has direct applications to optimization. Algorithms like the Newton-Raphson method [104] and the super-universal regularized Newton method [105] have  $GL(n, \mathbb{R})$  and  $O(n)$  covariance, respectively. One can build libraries with these symmetries and find data-driven accelerated optimization schemes fine-tuned to particular problems. Ongoing work with Nissa Watkins has shown that Newton’s method can be modified with additional tensors to accelerate convergence when near roots at the cost of lowering global convergence probability.

Many different sparse regression techniques have been proposed for model discovery, and they widely vary in speed and robustness. There is unfortunately no silver bullet for this problem. The SVD-based homogeneous regression techniques are robust for PDEs with coefficients across several orders of magnitude, but it is best to use a variety of methods and compare their results. There is no reason that good sparse regression techniques should be based on exactly minimizing a cost function. Sparsity promoting penalties like  $\ell_0$  and  $\ell_1$  regularization can introduce hyper-parameters that must be tuned carefully. Both the

sequential thresholding of SINDy and the SVD-based regression discussed here are instead iterative algorithms with halting conditions. The hyper-parameter  $\gamma$  of our SVD-based methods is dimensionless and should always be  $O(1)$ .

An open programming challenge remains to write software capable of automatically constructing libraries from a user supplied list of tensor fields, differential operators, and the symmetry group. Daniel Gurevich's implementation of SPIDER [57] automates library generation for  $\mathcal{L}_0$  and  $\mathcal{L}_1$  assuming  $O(n)$  covariance. In my MATLAB code [57], the library is manually specified. This is inconvenient for libraries beyond  $O(20)$  terms, although it allows rapid prototyping of small libraries.

A second challenge is to extend SPIDER to consider nonlocal interactions in space and time. SPIDER could potentially learn symmetry-covariant kernels if a basis is provided. The choice of basis for kernels is quite arbitrary, and libraries sizes would greatly increase. Furthermore, evaluating such kernels in weak form would require numerical integration in large dimensions. The MATLAB implementation of SPIDER [106] has computed six-dimensional integrals for the study of conformal invariance in electromagnetism. Regression algorithms must be tuned to carefully handle these large libraries.

# **Appendices**

## APPENDIX A

### SPIDER CODE

The most significant deliverable of my work has been MATLAB code [106] for carrying out the SPIDER algorithm. An independent Python implementation has been written by Daniel Gurevich [57]. While the construction of libraries is left as a job for the physicist, picking subdomains, efficient weak-form evaluations, and sparse regression have all been packaged. I would like to thank Jason Marfey, Daiyue Sun, and Carlos Silva Filho, Daniel Gurevich, Patrick Reinbold, and Roman Grigoriev for their direct and indirect contributions to this code. The code is available at [github.com/mgolden30/SPIDER](https://github.com/mgolden30/SPIDER) with some example implementations for an experimental 2D flow [36] and a 2+1D simulation of homogeneous quantum electrodynamic turbulence.

Since any code-specific tutorial will quickly become out of date, I will provide a conceptual overview. Each tensor library (scalar, vector, antisymmetric rank-2, ...) is its own MATLAB script. Running this script will fill memory with many objects, the important ones being

- $G$  - the integrated library matrix used in model discovery.
- labels - a cell of strings with the LaTeX representations of each library term. It is not important that the string be in LaTeX, this is just for convenience.
- scales - the vector of scales  $S_j$  associated with each library term  $\mathbf{f}_j$ . Since  $G$  has been divided by  $S_j$ , the coefficient vector obtained by  $Gc = 0$  will need multiplied by  $S_j$  to return to physical units.

One can always write a wrapper function to return these three objects to keep memory clean. The script `sparse_regression.m` then has several options for regression including combinatoric searches, GISR, and Reverse Regression.

Another important note is there is a primary weak-form integration function: `SPI-  
DER_integrate`. This function integrates a single scalar field with an arbitrary polynomial  
weight and handling arbitrary integration by parts. Integration for vectors and rank-2 ten-  
sors can be done by writing wrapper functions, which just calls the scalar integration routine  
on each component, and perhaps computing some weighted sum.

## REFERENCES

- [1] G. Scholz and F. Scholz, “First-order differential equations in chemistry,” *Chem-Texts*, vol. 1, no. 1, 2014.
- [2] Y. Kuang, *Delay differential equations with applications in population dynamics*. Academic Press, 1993.
- [3] T. G. F. James, *An introduction to fluid dynamics*. Clarendon Press, 1958.
- [4] M. Tsamparlis and A. Paliathanasis, “Symmetries of differential equations in cosmology,” *Symmetry*, vol. 10, no. 7, p. 233, 2018.
- [5] F. Brauer, C. Castillo-Chavez, and Z. Feng, *Mathematical models in epidemiology*. Springer, 2019.
- [6] J. Barrow-Green, *Poincaré and the three body problem*. American Mathematical Society, 1997.
- [7] F. Pretorius, “Simulation of binary black hole spacetimes with a harmonic evolution scheme,” *Classical and Quantum Gravity*, vol. 23, no. 16, 2006.
- [8] G. Pfister, H. Schmidt, K. A. Cliffe, and T. Mullin, “Bifurcation phenomena in taylor-couette flow in a very short annulus,” *Journal of Fluid Mechanics*, vol. 191, no. 1, p. 1, 1988.
- [9] I. A. Kuznetsov, *Elements of applied bifurcation theory*. Springer, 2023.
- [10] J. Ling, A. Kurzwski, and J. Templeton, “Reynolds averaged turbulence modelling using deep neural networks with embedded invariance,” *Journal of Fluid Mechanics*, vol. 807, pp. 155–166, 2016.
- [11] Z. Li, B. Dong, and Y. Wang, “Learning invariance preserving moment closure model for Boltzmann–BGK equation,” *Communications in Mathematics and Statistics*, vol. 11, no. 1, pp. 59–101, 2023.
- [12] M. Koch-Janusz and Z. Ringel, “Mutual information, neural networks and the renormalization group,” *Nature Physics*, vol. 14, no. 6, pp. 578–582, 2018.
- [13] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

- [14] T. Cohen and M. Welling, “Group equivariant convolutional networks,” in *International conference on machine learning*, PMLR, 2016, pp. 2990–2999.
- [15] P. Savarese, H. Silva, and M. Maire, “Winning the lottery with continuous sparsification,” *Advances in neural information processing systems*, vol. 33, pp. 11 380–11 390, 2020.
- [16] Y. Zhang, P. Tiño, A. Leonardis, and K. Tang, “A survey on neural network interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.
- [17] J. Bongard and H. Lipson, “Automated reverse engineering of nonlinear dynamical systems,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 9943–9948, 2007.
- [18] M. Schmidt and H. Lipson, “Distilling free-form natural laws from experimental data,” *Science*, vol. 324, no. 5923, pp. 81–85, 2009.
- [19] S.-M. Udrescu and M. Tegmark, “Ai feynman: A physics-inspired method for symbolic regression,” *Science Advances*, vol. 6, no. 16, eaay2631, 2020.
- [20] J. P. Crutchfield and B. McNamara, “Equations of motion from a data series ‘,’” *Complex systems*, vol. 1, pp. 417–452, 1987.
- [21] C. Yao and E. M. Bollt, “Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems,” *Physica D: Nonlinear Phenomena*, vol. 227, no. 1, pp. 78–99, 2007.
- [22] I.-C. Chou and E. O. Voit, “Recent developments in parameter estimation and structure identification of biochemical and genomic systems,” *Mathematical biosciences*, vol. 219, no. 2, pp. 57–83, 2009.
- [23] S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Discovering governing equations from data by sparse identification of nonlinear dynamical systems,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [24] A. A. Kaptanoglu, L. Zhang, Z. G. Nicolaou, U. Fasel, and S. L. Brunton, “Benchmarking sparse system identification with low-dimensional chaos,” *Nonlinear Dynamics*, pp. 1–22, 2023.
- [25] D. A. Messenger and D. M. Bortz, “Weak sindy for partial differential equations,” *Journal of Computational Physics*, vol. 443, p. 110 525, 2021.
- [26] ———, “Weak sindy: Galerkin-based data-driven model selection,” *Multiscale Modeling amp; Simulation*, vol. 19, no. 3, pp. 1474–1497, 2021.

- [27] ———, “Learning mean-field equations from particle data using wsindy,” *Physica D: Nonlinear Phenomena*, vol. 439, p. 133 406, 2022.
- [28] K. Kaheman, J. N. Kutz, and S. L. Brunton, “Sindy-pi: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, no. 2242, 2020.
- [29] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, “Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 478, no. 2260, 2022.
- [30] M. Bär, R. Hegger, and H. Kantz, “Fitting partial differential equations to space-time dynamics,” *Physical Review E*, vol. 59, no. 1, p. 337, 1999.
- [31] D. Xu and O. Khanmohamadi, “Spatiotemporal system reconstruction using fourier spectral operators and structure selection techniques,” *Chaos*, vol. 18, no. 4, p. 043 122, 2008.
- [32] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Data-driven discovery of partial differential equations,” *Science Advances*, vol. 3, no. 4, e1602614, 2017.
- [33] H. Schaeffer, “Learning partial differential equations via data discovery and sparse optimization,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2197, p. 20 160 446, 2017.
- [34] D. R. Gurevich, P. A. Reinbold, and R. O. Grigoriev, “Robust and optimal sparse regression for nonlinear pde models,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 10, p. 103 113, 2019.
- [35] P. A. Reinbold, D. R. Gurevich, and R. O. Grigoriev, “Using noisy or incomplete data to discover models of spatiotemporal dynamics,” *Physical Review E*, vol. 101, no. 1, p. 010 203, 2020.
- [36] P. A. Reinbold, L. M. Kageorge, M. F. Schatz, and R. O. Grigoriev, “Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression,” *Nature Communications*, vol. 12, 2021.
- [37] M. Golden, R. Grigoriev, J. Nambisan, and A. Fernandez-Nieves, “Physics-informed data-driven modeling of active nematics,” *Science Advances*, 2023.
- [38] P. Cvitanović, *Group theory*. Nordita, 1984.

- [39] R. Penrose, “Applications of negative dimensional tensors,” *Combinatorial mathematics and its applications*, vol. 1, pp. 221–244, 1971.
- [40] J.-H. Kim, M. S. Oh, and K.-Y. Kim, “Boosting vector calculus with the graphical notation,” *American journal of physics*, vol. 89, no. 2, pp. 200–209, 2021.
- [41] J. Butcher and G. Wanner, “Runge-kutta methods: Some historical notes,” *Applied Numerical Mathematics*, vol. 22, no. 1–3, pp. 113–151, 1996.
- [42] R. P. Chan and A. Y. Tsai, “On explicit two-derivative runge-kutta methods,” *Numerical Algorithms*, vol. 53, no. 2–3, pp. 171–194, 2009.
- [43] S. Weinberg, *Gravitation and cosmology: Principles and applications of the general theory of relativity*. Wiley India, 2017.
- [44] C. W. Misner, K. S. Thorne, J. A. Wheeler, and D. Kaiser, *Gravitation*. Princeton University Press, 2017.
- [45] R. M. Wald, *General relativity*. Univ. of Chicago Press, 2009.
- [46] E. A. Paschos, *Electroweak theory*. Cambridge University Press, 2023.
- [47] J. I. Kapusta, “Quantum chromodynamics at high temperature,” *Nuclear Physics B*, vol. 148, no. 3-4, pp. 461–498, 1979.
- [48] R. Penrose and W. Rindler, *Spinors and space-time*. Cambridge University Press, 1984.
- [49] G. Rosen, “Galilean invariance and the general covariance of nonrelativistic laws,” *American Journal of Physics*, vol. 40, no. 5, pp. 683–687, 1972.
- [50] J.-M. Lévy-Leblond, “Galilei group and galilean invariance,” in *Group theory and its applications*, Elsevier, 1971, pp. 221–299.
- [51] P. Havas, “Four-dimensional formulations of newtonian mechanics and their relation to the special and the general theory of relativity,” *Reviews of Modern Physics*, vol. 36, no. 4, p. 938, 1964.
- [52] M. Cariglia, “General theory of galilean gravity,” *Physical Review D*, vol. 98, no. 8, 2018.
- [53] G. Pinski, “Galilean tensor calculus,” *Journal of Mathematical Physics*, vol. 9, no. 11, pp. 1927–1930, 1968.

- [54] A. Actor, “Classical solutions of su (2) yang—mills theories,” *Reviews of Modern Physics*, vol. 51, no. 3, p. 461, 1979.
- [55] H. Mark and S. Fernbach, “Properties of matter under unusual conditions,” *New York: Interscience Publication*, 1969.
- [56] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Data-driven discovery of partial differential equations,” *Science Advances*, vol. 3, no. 4, 2017.
- [57] D. Gurevich, [https://github.com/sibirica/SPIDER\\_discrete](https://github.com/sibirica/SPIDER_discrete), 2023.
- [58] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
- [59] R. Tibshirani, “Regression shrinkage and selection via the lasso: A retrospective,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 73, no. 3, pp. 273–282, 2011.
- [60] P. R. Gill, A. Wang, and A. Molnar, “The in-crowd algorithm for fast basis pursuit denoising,” *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4595–4605, 2011.
- [61] E. T. Hale, W. Yin, and Y. Zhang, “Fixed-point continuation for  $\ell_1$ -minimization: Methodology and convergence,” *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1107–1130, 2008.
- [62] D. Malioutov, M. Cetin, and A. Willsky, “Homotopy continuation for sparse signal representation,” *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*,
- [63] P. Radchenko and G. M. James, “Variable inclusion and shrinkage algorithms,” *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1304–1315, 2008.
- [64] D. R. Gurevich, P. A. Reinbold, and R. O. Grigoriev, “Learning fluid physics from highly turbulent data using sparse physics-informed discovery of empirical relations (spider),” *arXiv preprint arXiv:2105.00048*, 2021.
- [65] J. Toner and Y. Tu, “Flocks, herds, and schools: A quantitative theory of flocking,” *Physical Review E*, vol. 58, no. 4, pp. 4828–4858, 1998.
- [66] J. Toner, Y. Tu, and S. Ramaswamy, “Hydrodynamics and phases of flocks,” *Annals of Physics*, vol. 318, no. 1, pp. 170–244, 2005.

- [67] S. Ramaswamy, “The mechanics and statistics of active matter,” *Annu. Rev. Condens. Matter Phys.*, vol. 1, no. 1, pp. 323–345, 2010.
- [68] M. C. Marchetti *et al.*, “Hydrodynamics of soft active matter,” *Rev. Mod. Phys.*, vol. 85, pp. 1143–1189, 3 2013.
- [69] F. Jülicher, S. W. Grill, and G. Salbreux, “Hydrodynamic theory of active matter,” *Reports on Progress in Physics*, vol. 81, no. 7, p. 076 601, 2018.
- [70] A. Doostmohammadi, J. Ignés-Mullol, J. M. Yeomans, and F. Sagués, “Active nematics,” *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.
- [71] V. Narayan, S. Ramaswamy, and N. Menon, “Long-lived giant number fluctuations in a swarming granular nematic,” *Science*, vol. 317, no. 5834, pp. 105–108, 2007.
- [72] N. Kumar, R. Zhang, J. J. De Pablo, and M. L. Gardel, “Tunable structure and dynamics of active liquid crystals,” *Science advances*, vol. 4, no. 10, eaat7779, 2018.
- [73] T. Sanchez, D. T. Chen, S. J. DeCamp, M. Heymann, and Z. Dogic, “Spontaneous motion in hierarchically assembled active matter,” *Nature*, vol. 491, no. 7424, pp. 431–434, 2012.
- [74] H. Li *et al.*, “Data-driven quantitative modeling of bacterial active nematics,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 3, pp. 777–785, 2018.
- [75] K.-T. Wu *et al.*, “Transition from turbulent to coherent flows in confined three-dimensional active fluids,” *Science*, vol. 355, no. 6331, 2017.
- [76] L. M. Lemma, S. J. DeCamp, Z. You, L. Giomi, and Z. Dogic, “Statistical properties of autonomous flows in 2D active nematics,” *Soft matter*, vol. 15, no. 15, pp. 3264–3272, 2019.
- [77] A. Opatthalage *et al.*, “Self-organized dynamics and the transition to turbulence of confined active nematics,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 11, pp. 4788–4797, 2019.
- [78] S. P. Thampi, R. Golestanian, and J. M. Yeomans, “Velocity correlations in an active nematic,” *Physical review letters*, vol. 111, no. 11, p. 118 101, 2013.
- [79] L. Giomi, M. J. Bowick, P. Mishra, R. Sknepnek, and M. Cristina Marchetti, “Defect dynamics in active nematics,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2029, p. 20 130 365, 2014.

- [80] S. P. Thampi, R. Golestanian, and J. M. Yeomans, “Instabilities and topological defects in active nematics,” *EPL (Europhysics Letters)*, vol. 105, no. 1, p. 18 001, 2014.
- [81] L. Giomi, “Geometry and topology of turbulence in active nematics,” *Physical Review X*, vol. 5, no. 3, p. 031 003, 2015.
- [82] S. P. Thampi, A. Doostmohammadi, R. Golestanian, and J. M. Yeomans, “Intrinsic free energy in active nematics,” *EPL (Europhysics Letters)*, vol. 112, no. 2, p. 28 004, 2015.
- [83] S. P. Thampi and J. M. Yeomans, “Active turbulence in active nematics,” *The European Physical Journal Special Topics*, vol. 225, no. 4, pp. 651–662, 2016.
- [84] R. Green, J. Toner, and V. Vitelli, “Geometry of thresholdless active flow in nematic microfluidics,” *Physical Review Fluids*, vol. 2, no. 10, p. 104 201, 2017.
- [85] A. Doostmohammadi, T. N. Shendruk, K. Thijssen, and J. M. Yeomans, “Onset of meso-scale turbulence in active nematics,” *Nature communications*, vol. 8, no. 1, pp. 1–7, 2017.
- [86] B. Martínez-Prat, J. Ignés-Mullol, J. Casademunt, and F. Sagués, “Selection mechanism at the onset of active turbulence,” *Nature physics*, vol. 15, no. 4, pp. 362–366, 2019.
- [87] R. Alert, J.-F. Joanny, and J. Casademunt, “Universal scaling of active nematic turbulence,” *Nature Physics*, vol. 16, no. 6, pp. 682–688, 2020.
- [88] J. L. Ericksen, “Conservation laws for liquid crystals,” *Transactions of the Society of Rheology*, vol. 5, no. 1, pp. 23–34, 1961.
- [89] F. M. Leslie, “Some constitutive equations for liquid crystals,” *Archive for Rational Mechanics and Analysis*, vol. 28, no. 4, pp. 265–283, 1968.
- [90] P.-G. De Gennes and J. Prost, *The physics of liquid crystals*, 83. Oxford university press, 1993.
- [91] A. N. Beris, B. J. Edwards, *et al.*, *Thermodynamics of flowing systems: with internal microstructure*, 36. Oxford University Press on Demand, 1994.
- [92] D. J. G. Pearce, P. W. Ellis, A. Fernandez-Nieves, and L. Giomi, “Geometrical control of active turbulence in curved topographies,” *Phys. Rev. Lett.*, vol. 122, p. 168 002, 16 2019.

- [93] J. Weickert, “Coherence-enhancing diffusion filtering,” *International Journal of Computer Vision*, vol. 31, no. 2/3, pp. 111–127, 1999.
- [94] P. W. Ellis, J. Nambisan, and A. Fernandez-Nieves, “Coherence-enhanced diffusion filtering applied to partially-ordered fluids,” *Molecular Physics*, vol. 118, no. e1725167, pp. 1–9, 2020.
- [95] J. V. Selinger, “Interpretation of saddle-splay and the oseen-frank free energy in liquid crystals,” *Liquid Crystals Reviews*, vol. 6, no. 2, pp. 129–142, 2018.
- [96] J. Ericksen, “Orientation induced by flow,” *Transactions of the Society of Rheology*, vol. 6, no. 1, pp. 275–291, 1962.
- [97] G. Ryskin, “One modification of the ericksen-leslie equations in a two-dimensional motion,” *Journal of non-newtonian fluid mechanics*, vol. 39, no. 2, pp. 207–210, 1991.
- [98] C. Joshi *et al.*, “Data-driven discovery of active nematic hydrodynamics,” *Physical review letters*, vol. 129, no. 25, p. 258 001, 2022.
- [99] 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*, vol. 75, no. 6, p. 1226, 1995.
- [100] J. Toner and Y. Tu, “Flocks, herds, and schools: A quantitative theory of flocking,” *Physical review E*, vol. 58, no. 4, p. 4828, 1998.
- [101] ———, “Long range order in a two-dimensional dynamical xy model: How birds fly together,” *Physical review letters*, vol. 75, no. 23, p. 4326, 1995.
- [102] C. Packard and D. M. Sussman, “Non-reciprocal forces and exceptional phase transitions in metric and topological flocks,” *arXiv preprint arXiv:2208.09461*, 2022.
- [103] E. P. Alves and F. Fiuza, “Data-driven discovery of reduced plasma physics models from fully kinetic simulations,” *Physical Review Research*, vol. 4, no. 3, p. 033 192, 2022.
- [104] T. J. Ypma, “Historical development of the newton–raphson method,” *SIAM review*, vol. 37, no. 4, pp. 531–551, 1995.
- [105] N. Doikov, K. Mishchenko, and Y. Nesterov, “Super-universal regularized newton method,” *arXiv preprint arXiv:2208.05888*, 2022.
- [106] M. Golden, <https://github.com/mgolden30/SPIDER>, 2023.