Research Statement

Zvi Rosen

December 14, 2017

My research interests lie at the intersection of **applied algebraic geometry** and **mathematical biology**. My introduction to mathematical biology was through algebraic matroids. As a Ph.D. student, I focused on the algebraic geometry and combinatorics found in coordinate projections of algebraic varieties. One application is chemical reaction network theory–experiments in which only some quantities are measurable can be thought of as coordinate projections. When the system is at steady-state, and modeled by polynomial ODE's, the resulting projections define an algebraic matroid. Performing a *valuation* of the matroid, i.e. computing the degrees of associated finite field extensions, can give insight into whcih experiments are "best" to infer the concentrations of every chemical species.

From there, I became interested in using the tools of algebraic geometry, discrete geometry, and combinatorics to answer other questions in biology. As I continued studying biology as a postdoctoral fellow, I began to expand my toolbox. In particular, I learned more about statistical tools for network inference, and I experimented with the incredibly powerful, if little understood, machinery of deep neural networks. While I am committed to exploring new questions in biology, I continue to research other applications of algebra, especially to the theory of tensors.

The remainder of this research statement will follow the structure alluded to above. First I will describe my projects in the intersection of algebra and biology; then, I will describe some projects in their symmetric difference.

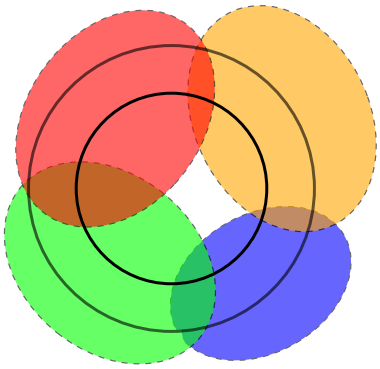
Contents

1 Algebraic Problems in Biology

1.1 Combinatorial Neural Codes

Experimentalists connected the brain of a freely moving rat to a collection of electrical sensors that measured neural activity in different areas of its hippocampus. They discovered something remarkable: there were neurons (later called *place cells*) that lit up precisely when the rat entered a certain region of its cage. As it paced around the cage, different sensors were ignited corresponding to each region; it was like a map inside of the rat's brain. This research won John O'Keefe part of the 2016 Nobel Prize in Medicine.

The brain processes its environment using input from the place cells – what part of the geometry of the space can we reconstruct from the firing patterns alone? With no constraints on place fields, this is obviously an ill-posed problem. The math steps into a much stronger role when we observe that the place cells in experimental settings corresponded to convex regions of the enclosure. For example, suppose that we have sensors in four place cells; we can represent on-off positions by binary strings of length four. Suppose we read the signals $\{1000, 1100, 0100, 0110, 0010, 0011, 0001, 1001\}$. This might lead us to guess that the rat occupies a circular track, with the place cells corresponding to overlapping segments.



A broader characterization of this endeavor is that it seeks to translate the combinatorial properties of neural firing patterns into geometric-topological properties of the place fields. My research has tackled this question in various forms:

Problem 1. What *combinatorial* properties of a code $\mathcal{C}$ correspond to what *topological* properties of a realization of $\mathcal{C}$?

Results from combinatorics, convex geometry, and algebraic topology are leveraged to obtain properties of these neural codes. Here is a sample of the results we have proved in three topological regimes:

Theorem 1.1. *([24, Proposition 2.1.2]) $\mathcal{C}$ is realizable in $\mathbb{R}^1$ as an arrangement of open intervals if and only if $\mathcal{C}$ is the column set of an "harmonious consecutive-ones matrix."*

In that paper, we also prove that this property can be checked in $O(n+k)$ time. In a large collaboration with participants in an AMS Math Research Community led by Carina Curto, our results focused on the simplicial complex of the code $\Delta(\mathcal{C})$, whose facets are the maximal codewords. We defined *local obstructions* which rest on applications of the Nerve Lemma to links of the complex.

Theorem 1.2. *([7, Theorem 1.3]) For each simplicial complex Δ, there is a unique minimal code $\mathcal{C}_{\min}(\Delta)$ with the following properties:*

1. *the simplicial complex of $\mathcal{C}_{\min}(\Delta)$ is Δ, and*
2. *for any code $\mathcal{C}$ with simplicial complex Δ, $\mathcal{C}$ has no local obstructions if and only if $\mathcal{C} \supseteq \mathcal{C}_{\min}(\Delta)$.*

In other words, given a set of maximal codewords, there is a corresponding set of "make-or-break" codewords that need to be present; if they are present, only non-local obstructions may exist. Finally, in ongoing work with Itskov & Kunin, we define $B\Delta(\mathcal{C})$, another simplicial complex associated to a neural code; this is especially useful for discussing codes associated to open half-space arrangements.

Theorem 1.3. *A neural code $\mathcal{C}$ is a generic hyperplane code only if $B\Delta(\mathcal{C})$ is shellable.*

My aim moving forward is to anchor the growing theory of combinatorial neural codes to other well-studied areas of combinatorics. In particular, under some restrictions, neural codes can be understood as the tope complexes of a class of oriented matroids. The theory of lopsided sets, antimatroids and convex geometries, posets that satisfy certain axioms of convex set arrangements, also has a lot to offer in the realm of neural codes. In my work, I will make these connections explicit, so that the respective communities can enrich each other's understanding.

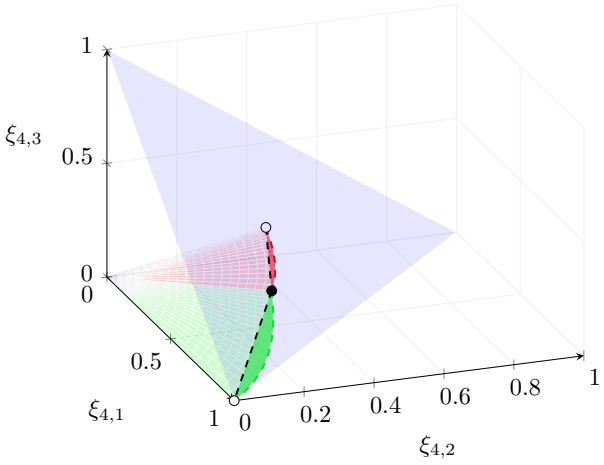
1.2 Population Genetics

Phylogenetics has been a topic of extensive study by the applied algebraic geometry community in the past decade. It has motivated important developments in toric and tropical geometry; see, for example, [21, 2]. A neighboring field that has been less penetrated by algebraic geometry is population genetics, the statistical study of genetic samples from a population. Together with Anand Bhaskar, Sebastien Roch, and Yun S. Song, we used the tools of algebraic geometry to answer questions from population genetics.

Fisher introduced the sample frequency spectrum (SFS) in 1930 [9] as a summary statistic for genomic data. Given a sample of n genomes from a population, the SFS is a vector of length $n-1$ counting the number of mutations common to k individuals in the sample, where k ranges from 1 to $n-1$. The SFS also appears in the context of modeling as a vector of probabilities; in particular, the k-th entry is the probability that a given mutation appears in k individuals.

Problem 2. Describe the geometry of the map from demographies, or popuation size histories, to sample frequency spectra.

Under the standard Wright-Fisher coalescent model with neutral selection (i.e. mutations do not lead to enhanced genetic fitness), the expected value of the SFS is polynomial in terms of piece-wise constant demographies. Previous work by [4] proved that restricting to demographies with a bounded number of epochs (i.e. pieces of a piecewise-defined function) allows this polynomial map to be injective. In our pre-print [23], we study the image of the polynomial map (depicted for $n = 4$ and 2 epochs at right). We were able to relate the set of neutral-selection spectra to both: (a) the secant variety of a rational curve, and (b) the cone over a Hadamard product of rational normal curves, leading to the following:



Theorem 1.4. *([23, Theorem 4.3]) Given a fixed sample size n, the selection-neutral sample frequency spectrum for any Lebesgue-measurable demography can be obtained from a piecewise-constant demography with at most $\kappa(n)$ epochs, where*

$$\left\lceil \frac{n}{2} \right\rceil \leq \kappa(n) \leq 2n-1.$$

A plethora of questions remain: For example, for fixed n is the set of all sample frequency spectra convex? This is only known to be true for $n \leq 4$. Using tools from real algebraic geometry, can we tighten the bounds on $\kappa(n)$, mentioned above? We plan to pursue each of these questions in future work. One direction is via Hadamard products of varieties, as recently explored in [5] and [10]. Other work we are pursuing in this realm seeks to generalize the results of [4] to more general models: in particular, 1) isolated populations with common ancestors 2) populations that were isolated but experienced migration events, and 3) populations with selective sweeps.

1.3 Chemical Reaction Network Theory

A chemical reaction network (CRN) can be defined as a directed graph, whose vertices are labeled with chemical complexes, and whose directed edges represent a chemical reaction that takes place. Each edge is labeled with a reaction rate, and the resulting network defines the system of ODE's governing the CRN dynamics. The recent proof of the global attractor conjecture by Craciun [6] demonstrates the power of sophisticated algebraic techniques to prove results about CRN dynamics. The majority of my work in this area focuses on the following problem:

Problem 3. What numerical properties of the steady-state solutions can be inferred from the combinatorial and algebraic properties of the ODE's?

In the paper [12], we used algebraic matroids to study experimental design for the Wnt Shuttle model. Given the steady-state equations coming from the ODE's on the 19 species, assuming generic parameters, and allowing the conserved quantities to vary freely, led to a steady-state variety of dimension 5. Close analysis of the algebraic matroid using `Macaulay2` [11] and `Bertini` [3] led to the following:

Proposition 1.5. *From the set of 19 species in the Wnt Shuttle model, there are 416 subsets of size 5 such that the conservation relations translate to a system with mixed volume 9 – the true number of expected solutions for the full system.*

This result allows us to transform our ideal so that solutions are robust to parameter error.

In [18], we studied the closely related topic of state-space models, translating key notions like *identifiability, observability*, and *distinguishability* into algebraic terms using Gröbner bases and algebraic matroids. My present work in CRN theory is a collaboration analyzing multistationarity and multistability in a new Wnt signaling model.

2 Statistics & Machine Learning for Biology

2.1 Deep Neural Networks for Translation Dynamics

Translation is the cellular process that takes mRNA as input, and constructs proteins that execute the functions of the cell. Roughly speaking, a ribosome attaches to a strand of mRNA, and then moves along the strand, attaching amino acids to a polypeptide which unfurls from the ribosome through the exit tunnel. Recent work by Khanh Dao Duc and Yun S. Song in [8] has revealed some compelling properties of translation dynamics. They found criteria with impressive predictive power for average translation rate at different positions in mRNA. One such criterion is the number of positively-charged codons in the [1:11] window and the number of negatively-charge codons in the [6:14] window of the exit tunnel. This result hints to more specific interactions that, in aggregate, produce these net effects.

Problem 4. Can deep neural networks identify more detailed electrostatic/hydrophobic effects on the rate of translation elongation?

The idea behind this project is to implement a deep neural network that can take as input the "big data" of an mRNA sequence, and, as output, predict the translation elongation rate of the ribosome at that position. Deep neural networks have been used in many biological contexts, including translation [22]. Our study takes a new approach by feeding input data motivated by physical-chemical concerns, and using rate approximations based on a more precise model of elongation.

In embarking on this project, I learned how to design networks with `TensorFlow` [1]. I plan to form new collaborations with biologists interested in experimenting with these tools.

2.2 Boolean Implication Networks for Single-Cell Data

The genome (i.e. DNA sequence) of an organism contains the full manual for cell function throughout the organism. Differential expression of genes across cell types allows blood cells to behave differently from neurons and liver cells. In the past, experimental techniques only allowed aggregate transcriptional data; one could take a large number of cells, blend them, and measure the number of RNA transcripts in the resulting stew. Recent experiments have facilitated the collection of transcriptional data from a single cell. This data gives us a window into regulatory networks within just one cell.

Problem 5. How can single-cell transcription data be used to infer gene-gene activation & inhibition networks?

The strategy we are pursuing involves Boolean implication networks [20]. We transform the cell-gene read count matrix (M_{ij} = number of reads of gene i in cell j) to a gene-gene implication score tensor (N_{ijk} = the magnitude of relationship k between gene i and gene j, where k can be i activates j, i inhibits j, etc.) We use tools of statistical inference to extract the most meaningful relationships in the tensor. Our data comes from the mouse embryonic stem cells in [14]. This work is part of a collaboration with Khanh Dao Duc and a undergraduate advisee Yutong Wang. We are developing software that can efficiently analyze any given single-cell data set.

3 Algebraic Statistics for Tensors

3.1 Matrix and Tensor Completion

Consider the joint probability density function of k independent discrete random variables; this corresponds to a tensor $T_{i_1,\ldots,i_k} = P(X_1 = i_1, \ldots, X_k = i_k)$ with entries in $[0,1]$ adding up to one. By independence, the tensor will factor as the outer product of the vectors $P(X_j = i_j)_{j\in[n_j]}$; therefore, it will have rank one. Our work was motivated by the completion problem: given a subset of the tensor entries, is it possible to fill in the remaining probabilities so that the tensor has rank 1 and has entries within the desired bounds? This is a special instance of the following problem:

Problem 6. Given a rank constraint and specified linear constraints, what subsets of entries of a tensor are *algebraically independent*? How do semi-algebraic constraints affect the problem?

The first results we obtained for this problem were for the special case of matrices. In my first paper on the topic with Kaie Kubjas, we obtained a nice description for the semi-algebraic constraints in terms of the combinatorics of the set of entries. We can consider matrix entries $\{M_{i,j}\}$ as edges in a bipartite graph with row vertices and column vertices; then sets of entries are like subgraphs of $K_{m,n}$.

Theorem 3.1. *([16, Theorem 4]) The projection of all independence matrices onto a forest with n trees has boundary given by coordinate hyperplanes and a hypersurface $\sum_{i=1}^{n}\sqrt{b_i} = 1$, where each b_i is a rational function in the entries of a single connected component.*

The situation for tensors is predictably more complicated. Still, the case of diagonal partial tensors (where all coordinates of all given entries are distinct) is relatively simple, and we proved the following result in the same paper:

Theorem 3.2. *([16, Theorem 7]) Let $a_1, \ldots, a_k$ be the entries in the diagonal of an order-d tensor. Then the diagonal partial tensor is completable to a rank-1 probability tensor if and only if*

$$\sum_{i=1}^{k} a_i^{1/d} \leq 1.$$

In a follow-up work with Kahle and Kummer [13], we answered the question for general partial tensors, and gave polynomial inequalities equivalent to the algebraic inequalities above.

3.2 Tensor Eigenvalues

The study of tensor eigenvectors has been a dynamic area of research since they were independently developed by Lim [17] and Qi [19] in 2005. Certain definitions of tensor eigenvectors and eigenvalues have gained popularity, particularly Qi's Z-eigenvalue. However, the eigenpairs defined in [15] have been relatively less studied, despite having a very natural definition. In joint work with D. Pham, M. Wang, and Y.S. Song, we address the following problem:

Problem 7. Does there exist a real symmetric tensor such that the maximal eigenvalue on the complex d-sphere is different from the maximal eigenvalue on the real d-sphere?

We demonstrate, in particular, that such a tensor does not exist in orders ≤ 4. We also show that a tensor with this property will have decomposition with tensors that have inner product of varying signs. This work is in progress, and we hope to answer the question completely.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Federico Ardila and Caroline J Klivans. The Bergman complex of a matroid and phylogenetic trees. *Journal of Combinatorial Theory, Series B*, 96(1):38–49, 2006.

[3] Daniel J Bates, Jonathan D Hauenstein, Andrew J Sommese, and Charles W Wampler. Bertini: Software for numerical algebraic geometry (2006). *Software available at http://bertini. nd. edu.*

[4] Anand Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals ofStatistics*, 42(6):2469, 2014.

[5] Cristiano Bocci, Enrico Carlini, and Joe Kileel. Hadamard products of linear spaces. *J. Algebra*, 448:595–617, 2016.

[6] Gheorghe Craciun. Toric differential inclusions and a proof of the global attractor conjecture. *arXiv preprint arXiv:1501.02860*, 2015.

[7] Carina Curto, Elizabeth Gross, Jack Jeffries, Katherine Morrison, Mohamed Omar, **Zvi Rosen**, Anne Shiu, and Nora Youngs. What makes a neural code convex? *SIAM Journal on Applied Algebra and Geometry*, 1(1):222–238, 2017.

[8] Khanh Dao Duc and Yun S Song. Identification and quantitative analysis of the major determinants of translation elongation rate variation. *bioRxiv*, page 090837, 2017.

[9] R. A. Fisher. The distribution of gene ratios for rare mutations. In *Proc. R. Soc. Edinb*, volume 50, pages 205–220, 1930.

[10] Netanel Friedenberg, Alessandro Oneto, and Robert L Williams. Minkowski sums and hadamard products of algebraic varieties. *arXiv preprint arXiv:1701.03191*, 2017.

[11] Daniel R Grayson and Michael E Stillman. Macaulay 2, a software system for research in algebraic geometry, 2002.

[12] Elizabeth Gross, Heather A Harrington, **Zvi Rosen**, and Bernd Sturmfels. Algebraic systems biology: a case study for the Wnt pathway. *Bulletin of Mathematical Biology*, 78(1):21–51, 2016.

[13] Thomas Kahle, Kaie Kubjas, Mario Kummer, and **Zvi Rosen**. The geometry of rank-one tensor completion. *SIAM Journal on Applied Algebra and Geometry*, 1(1):200–221, 2017.

[14] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[15] Tamara G Kolda and Jackson R Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

[16] Kaie Kubjas and **Zvi Rosen**. Matrix completion for the independence model. *Journal of Algebraic Statistics*, 8(1), 2017.

[17] Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 129–132. IEEE, 2005.

[18] Nicolette Meshkat, **Zvi Rosen**, and Seth Sullivant. Algebraic tools for the analysis of state space models. *arXiv preprint arXiv:1609.07985. To appear in Proceedings of Mathematical Society of Japan, 2015 Summer Institute on Grobner bases*, 2016.

[19] Liqun Qi. Eigenvalues of a real supersymmetric tensor. *Journal of Symbolic Computation*, 40(6):1302–1324, 2005.

[20] Debashis Sahoo, David L Dill, Andrew J Gentles, Robert Tibshirani, and Sylvia K Plevritis. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biology*, 9(10):R157, 2008.

[21] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(2):204–228, 2005.

[22] Sai Zhang, Hailin Hu, Jingtian Zhou, Xuan He, Tao Jiang, and Jianyang Zeng. Rose: a deep learning based framework for predicting ribosome stalling. *bioRxiv*, page 067108, 2016.

[23] **Zvi Rosen**, Anand Bhaskar, Sebastien Roch, and Yun S Song. Geometry of the sample frequency spectrum and the perils of demographic inference. *bioRxiv*, 2017.

[24] **Zvi Rosen** and Yan X Zhang. Convex neural codes in dimension 1. *arXiv preprint arXiv:1702.06907*, 2017.