Boston Collge

Morrissey College of Arts and Sciences Graduate School

Ph.D. Thesis

Statistical Mechanics of Microbiomes

Wenping Cui

Supervisor:
Prof. Pankaj Mehta

Thesis Committee:
Prof. Ziqiang Wang
Prof. Kevin Bedell
Prof. Ying Ran

submitted to the Faculty of the department of physics in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 2021

Statistical Mechanics of Microbiomes

Wenping Cui

Advisor: Prof. Pankaj Mehta

Nature has revealed an astounding degree of phylogenetic and physiological diversity in natural environments – especially in the microbial world. Microbial communities are incredibly diverse, ranging from 500-1000 species in human guts to over 10^3 species in marine ecosystems. Historically, theoretical ecologists have devoted considerable effort to analyzing ecosystems consisting of a few species. However, analytical approaches and theoretical insights derived from small ecosystems consisting of a few species may not scale up to diverse ecosystems. Understanding such large complex ecosystems poses fundamental challenges to current theories and analytical approaches for modeling and understanding the microbial world. One promising approach for tackling this challenge that I develop in my thesis is to adapt and expand ideas from statistical mechanics to theoretical ecology. Statistical mechanics has helped us to understand how collective behaviors emerge from the interaction of many individual components. In this thesis, I present a unified theoretical framework for understanding complex ecosystems based on statistical mechanics, random matrix theories and convex optimization. My thesis work has three key aspects: modeling, simulations, and theories.

Modeling: Classical ecological models often focus on predator-prey relationships. However, this is not the norm in the microbial world. Unlike most macroscopic organisms, microbes relie on consuming and producing small organic molecules for energy and reproduction. In this thesis, we develop a new Microbial Consumer Resource Model that take into account these types of metabolic cross-feeding interactions. We demonstrate that this model can qualitatively reproduce and explain statistical patterns observed in large survey data, including Earth Microbiome Project and the Human Microbiome Project.

Simulations: Computational simulations are essential in theoretical ecology. Complex ecological models often involve ordinary differential equations(ODE) containing hundreds to thousands of interacting variables. Typical ODE solvers are based on numerical integration methods, which are both time and resource intensive. To overcome this bottleneck, we derived a surprising duality between constrained convex optimization and generalized consumer-resource models describing ecological dynamics. This allows us to develop a fast algorithm to solve the steady state of complex ecological models. This improves computational performance by between 2-3 orders of magnitude compared to direct numerical integration of the corresponding ODEs.

Theories: Few theoretical approaches allow for the analytic study of communities containing a large number of species. Recently, there has been considerable interest in the idea that ecosystems can be thought of as a type of disordered systems. This mapping suggests that understanding community coexistence patterns is actually a problem in "spin glass" physics. This has motivated physicists to use insights from spin glass theory to uncover the universal features of complex ecosystems. In this thesis, I use and extend the cavity method, originally developed in spin glass theories, to answer fundamental ecological questions regarding the stability, diversity, and robustness of ecosystems. I use the cavity method to derive new species backing bounds and uncover novel phase transitions to typicality.

"The way was long, and wrapped in gloom did seem,
As I urged on to seek my vanished dream."

Qu Yuan

Acknowledgements

Foremost, my deepest gratitude is to my supervisor, Professor Pankaj Mehta, who proposed the basic idea of this thesis. Pankaj is the best teacher I have ever met. When I joined the group in 2016, I knew nothing about biology and academic research. He guided me into the field of biophysics and shaped my taste in science. Without his supervision and constant help, this dissertation would not have been possible. I appreciate all his contributions of time, ideas, and funding to make my research experience productive and stimulating.

My colleague, Robert Marsland III has always been there to listen and give advice. I am deeply grateful for his help in sorting out the technical details of my work. I also learned a lot from him about how to organize research projects. My collaborator, Marin Bukov taught me many coding skills. Even though all our meetings were on Skype, Marin gave me tremendous support in research and postdoc applications. I am also deeply thankful to my old friend, Mingda Li, who inspired me to come to United States from Germany and helped me a lot in the early stage of my Ph.D. study and my postdoc application. His encouragement and passion for research let me stick to academics. I also would like to thank Zhenyu Liao, who enlightened me the first glance of random matrix theories and answer my questions patiently. Professor Ziqiang Wang and Jane Carter kindly encouraged me to overcome the difficulties at the beginning of my graduate study. I would never forget their help during my studies at Boston College.

My friends, Tailin Wu, Guangwei Si, Hong Pan, and Tianchi Chen, especially my squash partner, He Zhao, have helped me stay sane through a difficult time. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship, and I sincerely appreciate their belief in me.

My thanks go to all my colleagues and friends: Ashish Bino George, Alex Golden, Jason Rocks, Alex Day, Ching-Hao Wang, Despina Bokios, Anita Gupta, Kelly Capri, Robyn Kinch, Thompson Scott-Ludwig, Zheng Ren, Tong Yang, Yun Peng, Shenghan Jiang, Kun Jiang, Mengliang Yao, Chaobin Yang, Wei Zhang, Xu Yang, Shang Gao, and all the staff members in the Physics department of Boston College. I will always remember all the fun we have had in the last six years.

I would like to thank my parents, Yongfu Cui and Xiulan Lu, my sister, Zheng Cui, my brother-in-law, Yu Wang. I cannot express enough gratitude to Zheng Cui and Yu Wang, who take care of my parents and look after the whole family while I am abroad. I am grateful for the love of my family which continues to support and nourish me throughout my life.

Finally, I thank with love to my wife Yiqing Yan, who tolerated my absences, pique and impatience, accompanied, supported, encouraged and helped me through this agonizing period. When the pandemic started in China, knowing nothing about Covid 19, she risked her life and quarantined in Thailand alone for two weeks in order to come to US and accompany me through the difficult 2020. 2020 is the most memorable year in my life, when I married and she got pregnant. Right now I am watching our son Shizhi sleeping in the crib while writing this final paragraph at 12:43 a.m 03/29/2021. The thesis is dedicated to Shizhi, who is my inspiration to achieve greatness and makes me become a harder, stronger and better person. You are the best thing that has ever happened to me. Love you and your mom.

Contents

List of Figures

Chapter 1

Introduction to mathematical models in ecology

1.1 Background

One of the most stunning aspects of the natural world is the variety of life present in most environments. Ecosystems consisting of many species can exhibit numerous fascinating large-scale, collective phenomenons and perform critical functions in cycling of matter and energy on earth. This serves as major motivation for studying the general principles governing complex ecosystems.

Historically, theoretical ecologists have devoted considerable effort to analyzing ecosystems consisting of a few species. This was largely due to experimental limitations stemming from the difficulty of collecting large-scale ecological data. Even though the absolute number of a specific species in an ecosystems can be huge, individuals are often distributes sparsely over a large wild area making it inefficient to monitor multiple species with field surveys. Temporally, the breeding-cycle of experimental mammals and plants is often months or years, making it time-consuming it difficult to perform controlled experiments.

In contrast, microbial communities do not suffer from many of these technical limitations. Microbial community surveys using RNA sequencing technology have revealed an astounding degree of phylogenetic and physiological diversity. Species diversity estimates range from 500-1000 species in human guts [HGK+12b] to over 10^3 species in marine ecosystems [SCC+15]. Spatially, millions of bacterias can survive on a single dish. Temporally, the cell division circle is around an hour. Furthermore, growth conditions can be manipulated easily by choosing different temperatures, nutrients supplies,

and species pools. This makes microbial ecosystems an ideal experimental framework for studying complex ecosystems.

Microorganisms can be identified using a gene region in the ribosome (16S rRNA), making it possible to measure the relative abundance of microbes in a community using DNA sequencing [CLW+11, SWGV14]. Numerous microbial datasets have been generated with high resolution across numerous communities. However, understanding such the large amounts of data being generated by sequencing experiments presents some daunting challenges to current theories and analytical approaches in theoretical ecology.

1.2 Mathematical modeling in ecology

Mathematical models are necessary to understand ecological data quantitatively. In general, there are four classes of variables appearing in most ecological models:

1. species populations, which are also direct observables in the data,
2. interaction variables, describing how species interact with other species or environments,
3. species or environment variables, such as the species' birth, death rate, and environmental resource supply rate,
4. dynamical variables, such as time and space.

In this thesis, we do not consider spatial processes and restrict ourselves to well-mixed populations. We also focus primarily of the steady-state dynamics of these models. For these reasons, the models presented in this thesis are restricted to the first three classes of variables discussed above.

1.2.1 Neutral theory and niche theory

There are two popular theories in ecology: niche theories that emphasize selection and species differences and neutral theory that emphasizes stochasticity and treats all species as identical. Both of these theories are commonly used to explain observed species coexistence patterns. We view these two perspectives as complementary rather than conflicting.

Neutral theory is inspired from the analogous theory in population genetics. In ecology, neutral theory deals with species within a the same trophic level, i.e., all species

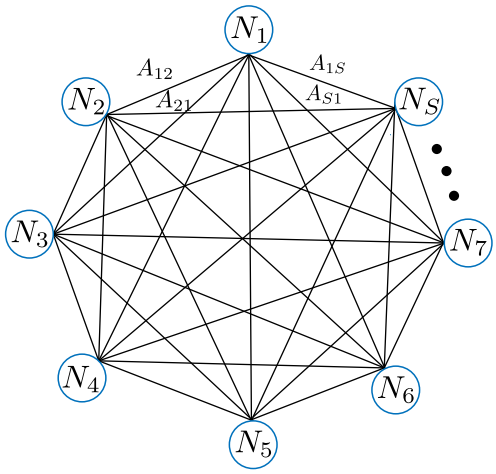

FIGURE 1.1: Schematic of Lotka–Volterra model

occupy the same niche and do not interact with each other and emphasizes the effect of stochastic drift and and migration on coexistence patterns. Stochastic processes are used extensively here. Let $\mathbf{N} = (N_1, N_2, N_3 \dots N_S)$ be a species abundance of an ecosystem with S species. Let $P(\mathbf{N}, t)$ be the probability of the state $\mathbf{N}$ at time t. Assuming that the stochastic dynamics are Markovian, the time evolution of $P(\mathbf{N}, t)$ can be expressed as a master equation:

$$\frac{\partial P(\mathbf{N}, t)}{\partial t} = \sum_{\mathbf{N}'} [T_{\mathbf{N}\mathbf{N}'} P(\mathbf{N}', t) - T_{\mathbf{N}'\mathbf{N}} P(\mathbf{N}, t)] \tag{1.1}$$

where $T_{\mathbf{N}\mathbf{N}'}$ and $T_{\mathbf{N}'\mathbf{N}}$ are the transition matrices. The behaviors of neutral theory can be investigated by studying mathematical properties of equation 1.1.

In contrast, niche theory is based on niche differentiation resulting from competing among species at the same trophic level. A fundamental result in niche theory is the competitive exclusion principle: each niche can only be occupied by at most one species. Niche theories mostly deal with purely deterministic processes, neglecting the stochastic effects emphasized in neutral theories. In niche theories, species interact with other species and the environments through fixed deterministic rules.

In this thesis I focus on the interactions between species and resources, i.e., niche theory, and refer readers interested in neutral theory to [ASG+16].

1.2.2 Lotka–Volterra model

The Lotka Volterra model describes the a model where species directly interact with each other (see Figure. 1.1). For an ecosystem with S species, the abundance N_i of species i is described by the ordinary differential equation

$$\frac{dN_i}{dt} = N_i(r_i - \sum_{i \neq j} A_{ij} N_j), \quad i = 1, 2, \dots S \tag{1.2}$$

where r_i is its intrinsic growth rate, and A_{ij} measures the interaction strength between population i and j. The factor N_i appearing outside the bracket ensures that when the species invade a new environment (i.e., all $N_j \ll 1$), they grow exponentially and that species abundances N_i never become negative. The second term in the bracket can be though of as the effective growth rate of species i. Notice that the presence of other species modifies the effective growth rate of species i, lowering it for competitive interactions ($A_{ij} > 0$) and raising it for a synergistic interaction ($A_{ij} < 0$).

From the physicist's perspective, we can analyze the Lotka–Volterra dynamics in terms of the first-order expansion near a stable fixed point, assuming such a fixed point exists. If the species' growth rate follows a general dynamics:

$$\frac{dN_i}{dt} = N_i g_i(\mathbf{N}), \tag{1.3}$$

we can expand the growth rate g_i to the first order around a fix point $\mathbf{N}^*$ to get

$$\frac{dN_i}{dt} = N_i \left[g_i(\mathbf{N}^*) + \sum_j \frac{\partial g_i}{\partial N_j}(N_j - N_j^*) + \mathcal{O}((N_j - N_j^*)^2) \right] \tag{1.4}$$

Relating equation 1.2 to equation 1.4 yields

$$r_i = g_i(\mathbf{N}^*) - \sum_j \frac{\partial g_i}{\partial N_j} N_j^*, \quad A_{ij} = -\frac{\partial g_i}{\partial N_j}. \tag{1.5}$$

Lotka–Volterra models can exhibit rich dynamical behaviors, even for a small ecosystem (see Figure 1.2). Let's consider an ecosystem consisting of two species,

$$\begin{cases} \frac{dN_1}{dt} = N_1(r_1 - A_{11}N_1 - A_{12}N_2) \\ \frac{dN_2}{dt} = N_2(r_2 - A_{21}N_1 - A_{22}N_2) \end{cases} \tag{1.6}$$

By solving $\frac{dN_1}{dt} = 0$ and $\frac{dN_2}{dt} = 0$, the steady state abundances can be written as

$$\bar{N_1} = \text{Max}\left[0, \frac{A_{22}r_1 - A_{12}r_2}{A_{11}A_{22} - A_{21}A_{12}}\right], \quad \bar{N_2} = \text{Max}\left[0, \frac{A_{11}r_2 - A_{21}r_1}{A_{11}A_{22} - A_{21}A_{12}}\right]. \tag{1.7}$$

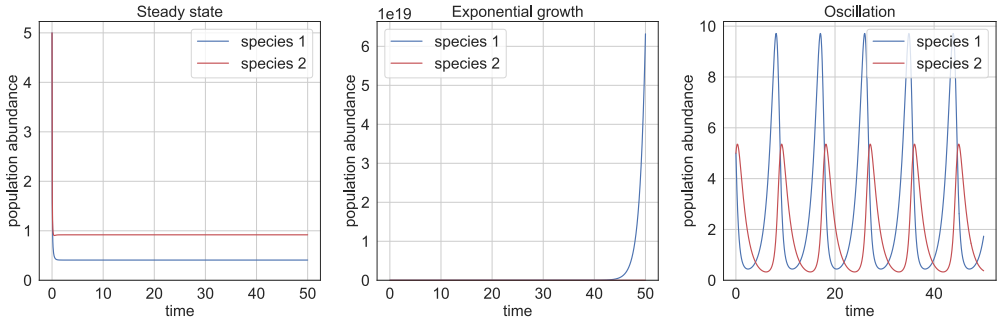

FIGURE 1.2: Different phases of Lotka-Volterra systems

However, such as steady-state may not exist or be stable. A trivial example of this if $A_{ij} = 0$ are zero and r_i is positive. In this case, the species abundance grows exponentially and will go to infinity after a long time. These equations can also exhibit periodic oscillation similar to harmonic motion. We can perturb equation around its fix point and only keep the linear terms [Mur07],

$$\begin{pmatrix} \frac{d\delta N_1}{dt} \\ \frac{d\delta N_2}{dt} \end{pmatrix} = \mathbf{J} \begin{pmatrix} \delta N_1 \\ \delta N_2 \end{pmatrix}, \quad \mathbf{J} = - \begin{pmatrix} \bar{N}_1 A_{11} & \bar{N}_1 A_{12} \\ \bar{N}_2 A_{21} & \bar{N}_2 A_{22} \end{pmatrix} \tag{1.8}$$

where $\delta N_i = N_i - \bar{N_i}^*$. If the eigenvalues of $\mathbf{J}$ are purely imaginary, we can expect the solutions in the neighborhood of the fixed point $(\bar{N}_1, \bar{N}_2)$ are periodic. We will return to these ideas when we analyze Lotka–Volterra models using methods from statistical physics in the limit where the number of species S becomes large.

1.2.3 MacArthur's consumer resource model

We now introduce another commonly used ecological model, MacArthur's consumer resource model (MCRM). In contrast to Lotka–Volterra model, the MCRM has no direct species-species interactions. Instead, species consume resources present in the ecosystem [ML67a] and species-species interactions emerge indirectly through competition for common resources. As shown in Figure 1.3, the MacArthur Consumer Resource Model consists of S species or consumers with abundances N_i $(i = 1...S)$ that can consume one of M substitutable resources with abundances R_α $(\alpha = 1...M)$, whose dynamics are described by the equations

$$\begin{aligned} \frac{dN_i}{dt} &= N_i(\sum_\beta C_{i\beta} R_\beta - m_i) \\ \frac{dR_\alpha}{dt} &= R_\alpha(K_\alpha - R_\alpha) - \sum_j N_j C_{j\alpha} R_\alpha. \end{aligned} \tag{1.9}$$

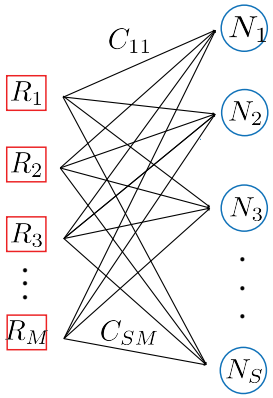

FIGURE 1.3: Schematic of MacArthur's consumer resource model

The consumption rate of species i for resource α is encoded by the entry $C_{i\alpha}$ in the $S \times M$ consumer preference matrix $\mathbf{C}$, K_α is the carrying capacity of resource α, and m_i is maintenance energy that encodes the minimum amount of energy that a species i must harvest from the environment in order to survive. $R_\alpha(K_\alpha - R_\alpha)$ is the resource supply dynamics, taken to be logistic growth in the original MCRM. In Chapter 6, we will discuss other forms of resource dynamics and how it affects the community properties.

When the system is in the steady state, some species and resources can vanish. We denote the numbers of surviving species and resources by S^* and M^*, respectively, and in general, at steady state we will have $S^* \leq M^*$. From an ecological view, we can interpret different types of resources as different niches. For M resources, at most, there exist M niches, resulting in M surviving species.

The Lotka-Volterra model can be derived from the MacArthur's consumer-resource model by assuming resource dynamics are much faster than species dynamics. Solving for the steady-state values of the non-extinct resources by setting the bottom equation in (1.9) equal to zero gives:

$$\bar{R}_\alpha = K_\alpha - \sum_i N_i C_{i\alpha} \tag{1.10}$$

Substituting this into the top equation in (1.9) gives:

$$\frac{dN_i}{dt} = N_i \left(\sum_{\alpha \in \mathbf{M}^*} C_{i\alpha} K_\alpha - m_i - \sum_j A_{ij} N_j \right) \tag{1.11}$$

where we have defined an interaction matrix $A_{ij} = \sum_{\alpha \in \mathbf{M}^*} \bar{C}_{i\alpha} C^T_{\alpha j}$ and $\mathbf{M}^*$ is the set of surviving resources. We can use this equation to solve for the steady-state (equilibrium)

abundances of non-extinct species, and arrive at the expression:

$$\bar{N}_i = \sum_{j \in \mathbf{S}^*} A_{ij}^{-1} \Big(\sum_{\alpha \in \mathbf{M}^*} C_{j\alpha} K_\alpha - m_j \Big)$$

where $\mathbf{S}^*$ is the set of surviving species. In terms of $\bar{N}_i$, the Lotka-Volterra equations become:

$$\frac{dN_i}{dt} = -\bar{N}_i \sum_j A_{ij} (N_j - \bar{N}_j) \tag{1.12}$$

In the future chapters, we will repeatedly return to variations of equation 1.9 and investigate its mathematical properties and ecological predictions for different ecological assumptions.

1.3 Mathematical modeling of microbial ecosystems

Microbial communities appear at every corner of our planet, from our own nutrient-rich guts to the remote depths of the ocean floor. The functional structure of these communities is highly variable, with functional traits often reflecting the environment in which the communities are found [TSM+17, HGK+12a]. A central goal of microbial community ecology is to understand how the effects of environments on diversity, stability, and functional structure [WAP+16]. And thus, it is important to build mathematical models to understand the mechanisms behind experimental phenomena.

However, the classical ecological models, based on niche competition, do not suit microbial communities. MacArthur's consumer resource model focuses on competition for resources. While it is true that bacteria compete for nutrients, they also often produce new resources in the form of metabolic byproducts. For this reason, the role of microbes is not limited to being a consumer, but also "producer" [GLB+18, HRD+14, ZS16]. The small molecules bacteria produce during metabolism always leak out into the environment and provide nutrients for other species. Crossfeeding helps change environments and shape species composition. In this thesis, we show that this little difference is one of the essential distinctions between classical and microbial ecology and can lead to dramatically different large-scale ecological properties.

1.3.1 Microbial consumer resource Model

Our starting point is to extend MacArthur's Consumer Resource Model to include crossfeeding interaction by considering the energy flux, the exchange, and consumption of

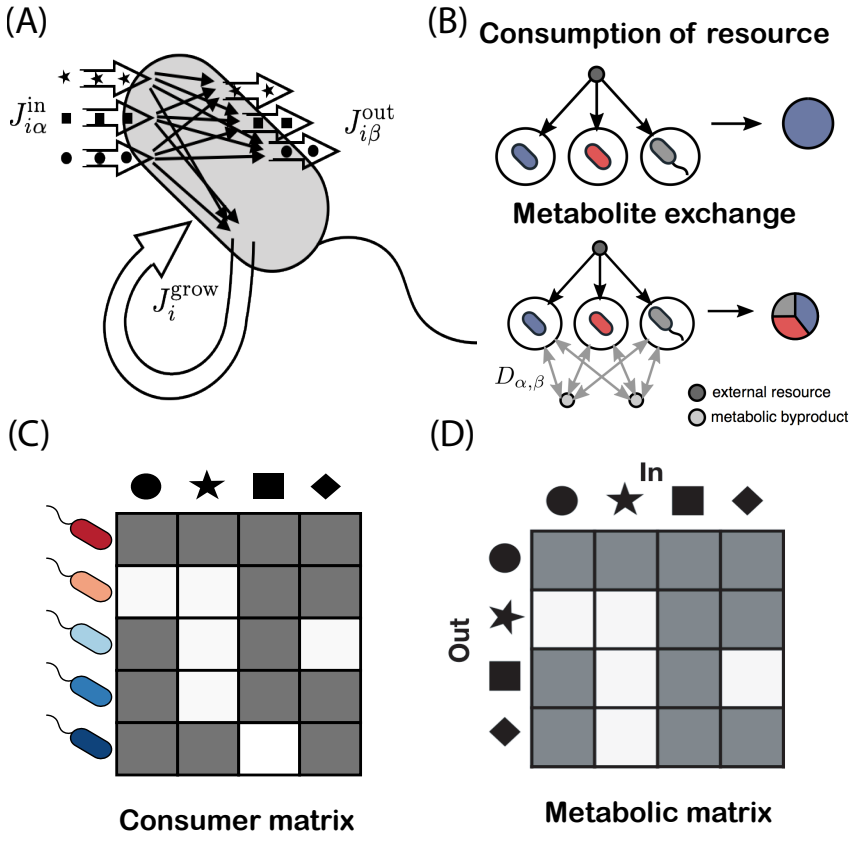

FIGURE 1.4: Schematic of (A) microbe-mediated energy fluxes in Microbial Consumer Resource Model; (B) Consumption of resource and metabolite exchange; (C) consumer matrix; (D) metabolic matrix. Made by Robert Marsland III in [MCM20]

metabolites (see Figure 1.4).

Just like in the original MacArthur's consumer resource model, the rate at which species i harvests energy from resource α depends on the resource concentration R_α and species' consumer preferences $C_{i\alpha}$. This is encoded in the input flux

$$J_{i\alpha}^{\text{in}} = C_{i\alpha} R_\alpha. \tag{1.13}$$

A core assumption of our model is that a species can not fully utilize the whole input energetic flux and releases some this energy back into the environment as leaked byproducts. We assume that a fraction $l_\alpha (< 1)$ of the input energy $J_{i\alpha}^{\text{in}}$ returns to the environment so that the power available to the cell for sustaining growth is

$$J_i^{\text{grow}} = \sum_\alpha (1 - l_\alpha) J_{i\alpha}^{\text{in}}. \tag{1.14}$$

The time-evolution of the species abundance N_i can be described with the equation

$$\frac{dN_i}{dt} = N_i \left(J_i^{\text{grow}} - m_i\right), \tag{1.15}$$

where m_i is the maintenance cost for species i.

The leaked energy flux $J_i^{\text{out}} = \sum_\alpha l_\alpha J_{i\alpha}^{\text{in}}$ from each cell of species i is partitioned among the M possible resource types via the biochemical pathways operating within the cell. We assume that all species share the same metabolism, encoded in a transformation matrix $D_{\beta\alpha}$. Each element of $D_{\beta\alpha}$ specifies the fraction of leaked energy from resource α that is released in the form of resource β. To enforce energy conservation, we have $\sum_\beta D_{\beta\alpha} = 1$. Thus, the outgoing energy flux contained in metabolite β is given by

$$J_{i\beta}^{\text{out}} = \sum_\alpha D_{\beta\alpha} l_\alpha J_{i\alpha}^{\text{in}}. \tag{1.16}$$

The resouce dynamics depends on the incoming and outgoing energy fluxes through the equations

$$\frac{dR_\alpha}{dt} = h_\alpha(R_\alpha) + \sum_j N_j (J_{j\alpha}^{\text{out}} - J_{j\alpha}^{\text{in}}), \tag{1.17}$$

where $h_\alpha(R_\alpha)$ is the resources dynamics in the absence of any microbes. In MacArthur's consumer resource model, $h_\alpha = R_\alpha(K_\alpha - R_\alpha)$ is assumed to be logistic. While such resource dynamics is reasonable for biotic resources, abiotic resources, such as minerals and small molecules cannot self-replicate and are usually supplied externally to the ecosystem. A simple way to model this scenario is by using linearized resource dynamics of the form,

$$h_\alpha(R_\alpha) = \kappa_\alpha - \tau_\alpha^{-1} R_\alpha. \tag{1.18}$$

where τ_α is the degredation rate of resource α. In many of the experiments that motivate our work, microbial communities are grown in minimum synthetic environments with a single externally supplied resource $\alpha = 0$. To model such experiments, all κ_α are set to zero except κ_0. These equations for N_i and R_α, along with the expressions for $J_{i\alpha}^{\text{in}}$ and $J_{i\alpha}^{\text{out}}$, completely specify the ecological dynamics of the model:

$$\begin{aligned} \frac{dN_i}{dt} &= N_i \left[\sum_\alpha (1 - l_\alpha) C_{i\alpha} R_\alpha - m_i\right], \\ \frac{dR_\alpha}{dt} &= \kappa_\alpha - \tau_\alpha^{-1} R_\alpha - \sum_i N_i C_{i\alpha} R_\alpha + \sum_{i,\beta} D_{\alpha\beta} l_\beta N_i C_{i\beta} R_\beta. \end{aligned} \tag{1.19}$$

An immediate technical problem that arises is to understand how to solve the above

dynamics when the number of species and resources becomes extremely large. I discuss this in Chapter 2.

Chapter 2

Numerical simulations of complex ecosystems

Computational simulations are essential in theoretical ecology. Complex ecological models always involve ordinary differential equations(ODE) containing hundreds to thousands of interacting variables. Typical ODE solvers are based on Runge–Kutta methods, which are both time and resource consuming, motivating us to develop fast simulation algorithms for complex ecological models. In this Chapter, we show a surprising duality between constrained optimization with inequality constraints and generalized consumer-resource models describing ecological dynamics [MCWMI19, MICM20], allowing us to develop a new Python package for simulating complex ecosystems [MCGM20]. Using this duality to solve for steady-state dynamics speeds-up performance by between 2-3 orders compared to direct numerical integration of the corresponding ODEs. Employing this package, we can reproduce large-scale patterns in microbial biodiversity from the Human Microbiome Project, Earth Microbiome Project, and similar surveys [MCM20].

2.1 Duality between constrained convex optimization and ecological dynamics

Optimization is an important problem for numerous disciplines, including physics, computer science, information theory, machine learning, and operations research [BV04, Ber99, MM09]. Many optimization problems are amenable to analysis using techniques from the statistical physics of disordered systems [Zde09, MPZ02, MM11]. Over the last few years, similar methods have been used to study community assembly and ecological dynamics suggesting a deep connection between ecological models of community

assembly and optimization [FM14, KS15, DFM16, Bun17, ABM18b, BABL18, BBC18b, TM17, MICG+19].Yet, the exact relationship between these two fields remains unclear.

Here, we show that constrained optimization problems with inequality constraints are naturally dual to an ecological dynamical system describing a generalized consumer resource model [ML67a, Mac70, Che90a]. As an illustration of this duality, we start a particular important and commonly encountered constrained optimization problem: quadratic programming (QP) [BV04]. In QP, the goal is to minimize a quadratic objective function subject to inequality constraints. We show that QP is dual to one of the most famous models of ecological dynamics, MacArthur's Consumer Resource Model (MCRM) introduced in Chapter 1, a system of ordinary differential equations describing how species compete for a pool of common resources [ML67a, Mac70, Che90a]. We also show that the Lagrangian dual of QP has a natural description in terms of generalized Lotka-Volterra equations that can be derived from the MCRM in the limit of fast resource dynamics. Later, we will generalize our results to other consumer-resource model dynamics.

2.1.1 Optimization as ecological dynamics

Consider an optimization problem of the form

$$\begin{aligned} \underset{\mathbf{R}}{\text{minimize}} \quad & f(\mathbf{R}) \\ \text{subject to} \quad & g_i(\mathbf{R}) \leq 0,\ i = 1, \ldots, S. \\ & R_\alpha \geq 0,\ \alpha = 1, \ldots, M. \end{aligned} \tag{2.1}$$

where the variables being optimized $\mathbf{R} = (R_1, R_2, \ldots, R_M)$ are constrained to be non-negative. We can introduce a 'generalized' Lagrange multiplier λ_i for each of the S inequality constraints in our optimization problem. In terms of the λ_i, we can write a set of conditions collectively known as the Karush-Kuhn-Tucker (KKT) conditions that must be satisfied at any local optimum $\mathbf{R}_{min}$ of our problem [BV04, Ber99, Bis06]. We note that for this reason, in the optimization literature the λ_i are often called KKT-multipliers rather than Lagrange multipliers. The KKT conditions are:

$$\begin{aligned} \textit{Stationarity:} \quad & \boldsymbol{\nabla}_{\mathbf{R}} f(\mathbf{R}_{\min}) + \sum_j \lambda_j \boldsymbol{\nabla}_{\mathbf{R}} \quad g_j(\mathbf{R}_{\min}) = 0 \\ \textit{Primal feasibility:} \quad & g_i(\mathbf{R}_{\min}) \leq 0 \\ \textit{Dual feasibility:} \quad & \lambda_i \geq 0 \\ \textit{Complementary slackness:} \quad & \lambda_i g_i(\mathbf{R}_{\min}) = 0, \end{aligned}$$

where the last three conditions must hold for all $i = 1, \ldots, M$. The KKT conditions have a straightforward and intuitive explanation. At the optimum $\mathbf{R}_{\min}$, either $g_i(\mathbf{R}_{\min}) = 0$ and the constraint is active $\lambda_i \geq 0$, or $g_i(\mathbf{R}_{\min}) \leq 0$ and the constraint is inactive $\lambda_i = 0$. In our problem, the KKT conditions must be supplemented with the additional requirement of positivity $R_\alpha \geq 0$.

One can easily show that the four KKT conditions and positivity are also satisfied by the steady states of the following set of differential equations restricted to the space $\lambda_i, R_\alpha \geq 0$:

$$\frac{d\lambda_i}{dt} = \lambda_i g_i(\mathbf{R}), \quad \frac{dR_\alpha}{dt} = [-\partial_{R_\alpha} f(\mathbf{R}) - \sum_j \lambda_j \partial_{R_\alpha} g_j(\mathbf{R})] R_\alpha. \tag{2.2}$$

The first of these equations just describes exponential growth of a "species" i with a resource-dependent "growth rate" $g_i(\mathbf{R})$. Species with $g_i(\mathbf{R}_{\min}) \leq 0$ correspond to constraints that are inactive and go extinct in the ecosystem (i.e, $\lambda_{i\,\min} = 0$), whereas species with $g_i(\mathbf{R}_{\min}) = 0$ survive at steady state and correspond to active constraints with $\lambda_{i\,\min} \neq 0$ (see Fig. 2.1 for a simple two-dimensional example). The second equation in (2.2) performs a "generalized gradient descent" on the optimization function $f(\mathbf{R}) + \sum_j \lambda_j g_j(\mathbf{R})$ (note the extra factor of R_α in our dynamics compared to the usual gradient descent equations). In the context of ecology, these equations describe the dynamics of a set of resources $\{R_\alpha\}$ produced at a rate $-\partial_{R_\alpha} f(\mathbf{R}) R_\alpha$ and consumed by individuals of species j at a rate $\lambda_j \partial_\alpha g_j(\mathbf{R}) R_\alpha$.

2.1.2 Ecological duals of Quadratic Programming (QP)

The optimization function of QP is quadratic, $f(\mathbf{R}) = \frac{1}{2}\mathbf{R}^T Q \mathbf{R} + \mathbf{b}^T \mathbf{R}$, with Q a positive semidefinite matrix, and linear inequality constraints. The positivity of Q guarantees that the problem is convex. By going to the eigenbasis of Q, we can always rewrite the QP problem as minimizing a square distance

$$\begin{aligned} \underset{\mathbf{R}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{R} - \mathbf{K}||^2 \\ \text{subject to} \quad & \sum_\alpha C_{i\alpha} R_\alpha \leq m_i, \; i = 1, \ldots, S. \\ & R_\alpha \geq 0, \; \alpha = 1, \ldots, M. \end{aligned} \tag{2.3}$$

Following (2.2), we introduce Lagrange (KKT) multipliers λ_i dual to each of the S constraints and Langrange KKT (multipliers) μ_α. that enforce positivity. Then, the

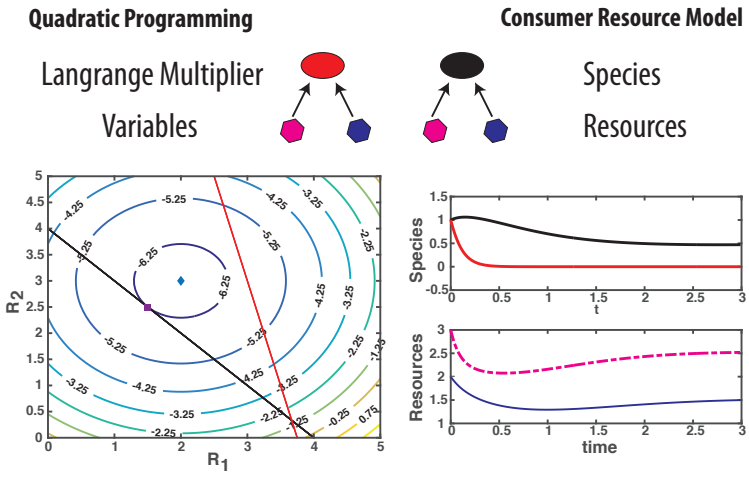

FIGURE 2.1: **Constrained optimization with inequality constraints is dual to an ecological dynamical system described by a generalized consumer resource model (MCRM)**. The variables to be optimized (hexagons) and Lagrange multipliers (ovals) are mapped to resources and species respectively. Species must consume resources to grow. (Bottom left) A quadratic programming (QP) problem with two inequality constraints where the unconstrained optimum differs from the constrained optimum. (Bottom right) Dynamics for MacArthur's Consumer Resource Model that is dual to this QP problem. The steady-state resource or species abundances correspond to the value of variables or Lagrange multipliers at the QP optimum. For this reason, species corresponding to inactive constraints go extinct. Made Pankaj Mehta in [MCWMI19]

function to be optimized is

$$\begin{aligned} &\underset{\lambda_j}{\text{maximize}} \quad \underset{\mathbf{R}_\alpha}{\text{minimize}} \frac{1}{2}\sum_\alpha (R_\alpha^2 - 2K_\alpha R_\alpha + K_\alpha^2) + \sum_{j,\alpha} \lambda_j (C_{j\alpha} R_\alpha - m_i) - \mu_\alpha R_\alpha \\ &\text{subject to} \qquad\qquad \lambda_j \geq 0 \; j = 1, \ldots, S \end{aligned} \tag{2.4}$$

We take the derivative with respect to R_α and note that

$$R_{\alpha *} = \max[0, K_\alpha - \sum_j C_{j\alpha}\lambda_j] \tag{2.5}$$

where we have used the KKT condition $\mu_\alpha R_{\alpha *} = 0$

Plugging this back into (2.4), we find that the function to be maximized with respect to the λ_i is

$$\sum_i \lambda_i [\kappa_i - \frac{1}{2}\sum_j A_{ij}\lambda_j] \tag{2.6}$$

with

$$\kappa_i = \sum_{\alpha, R_{\alpha *} \neq 0} K_\alpha C_{i\alpha} - m_i \tag{2.7}$$

and

$$A_{ij} = \sum_{\alpha, R_{\alpha *} \neq 0} C_{i\alpha} C_{j\alpha}. \tag{2.8}$$

We can construct the dual ecological model:

$$\begin{aligned} \frac{d\lambda_i}{dt} &= \lambda_i (\sum_{\alpha} C_{i\alpha} R_\alpha - m_i) \\ \frac{dR_\alpha}{dt} &= R_\alpha (K_\alpha - R_\alpha) - \sum_j \lambda_j C_{j\alpha} R_\alpha. \end{aligned} \tag{2.9}$$

The is the famous MacArthur Consumer Resource Model (MCRM), the same as equation 1.9, which was first introduced by MacArthur and Levins in their seminal papers [ML67b, Mac70] and has played an extremely important role in theoretical ecology [Che00, Til82a].

In optimization problems, one often works with the Lagrangian dual of an optimization problem. We show in Chapter 1 that the dual to equation 1.12 is just

$$\begin{aligned} \underset{\lambda_i}{\text{maximize}} \quad & \sum_i \lambda_i [\kappa_i - \frac{1}{2} \sum_j A_{ij} \lambda_j] \\ \text{subject to} \quad & \lambda_i \geq 0, \end{aligned} \tag{2.10}$$

the sum restricted to α for which $R_{\alpha\,\min} \neq 0$. It is once again straightforward to check that the local minima of this problem are in one-to-one correspondence with steady states of the Generalized Lotka-Volterra Equations (GLVs) of the form:

$$\frac{d\lambda_i}{dt} = \lambda_i (\kappa_i - \sum_j A_{ij} \lambda_j) \tag{2.11}$$

As with the primal problem, the species in the GLV have a natural interpretation as Lagrange multipliers enforcing inequality constraints. This GLV can also be directly obtained from the MCRM in equation 2.9) in the limit where the resource dynamics are extremely fast by setting $\frac{dR_\alpha}{dt} = 0$ in the second equation and plugging in the steady-state resource abundances into the first equation [Mac70, Che90a]. This shows the Lagrangian dual of QP maps to a dynamical system described by a GLV – which itself can be derived from the MCRM which is the dynamical dual to the primal optimization problem!

2.2 Minimization Principle for generalized consumer resource models

This results derived in Section 2.1 are important for understanding the nature of steady states in ecological models. A key limitation of the derivation in the last section is that it is limited to cases where there is only one fixed point for the dynamics. This can be seen by noting that there exists a global Lyapunov function $f(\mathbf{R})$ for the dynamics. Here we show that optimization ideas presented in the last section can be extended to larger classes of population dynamics, suggesting that an optimization approach applies much more broadly than previously supposed [MICM20, TM17].

2.2.1 General derivation

Let's consider a general form of the consumer-resource model:

$$\begin{aligned} \frac{dN_i}{dt} &= N_i g_i(\mathbf{R}) \\ \frac{dR_\alpha}{dt} &= h_\alpha(\mathbf{R}) + \sum_i N_i q_{i\alpha}(\mathbf{R}). \end{aligned} \tag{2.12}$$

$g_i(\mathbf{R})$ is the growth rate. $\mathbf{q}_i(\mathbf{R})$ specifies the magnitude and direction of the resource abundance change induced by a single individual of the species [Til82b, Lei95]. The function $\mathbf{h}(\mathbf{R})$ encodes the externally supplied resource dynamics [Til82b, CL03].

Comparing equation 2.12 with equation 2.2, yields the following relation between consumer resource quantities and quantities appearing in the optimization problem:

$$q_{i\alpha} = -\frac{\partial g_i(\mathbf{R})}{\partial R_\alpha} R_\alpha, \quad h_\alpha = -R_\alpha \frac{\partial f(\mathbf{R})}{\partial R_\alpha}. \tag{2.13}$$

Surprisingly, the objective function in equation 2.3 of the corresponding optimization problem depends only on the function $\mathbf{h}(\mathbf{R})$, which characterizes the resource dynamcis in the absence of consumers, through the relation

$$f(\mathbf{R}) = -\sum_\alpha \int_{K_\alpha}^{R_\alpha} \frac{h_\alpha(\mathbf{x})}{x_\alpha} dx_\alpha. \tag{2.14}$$

Note that we are free to add a constant to $f(\mathbf{R})$ while still satisfying the conditions and can therefore always make $f(\mathbf{K}) = 0$ at its unconstrained minimum $\mathbf{K}$ (the carrying capacity of the resources without any consumers).

With the help of equation 2.14, we can obtain the objective function for different consumer resource models through a direct integral. He we illustrate this for two simple

variants of the MacArthur's consumer resource model. For the MacArthur's consumer resource model with logistic growth, $h_\alpha(\mathbf{R}) = R_\alpha(K_\alpha - R_\alpha)$, the objective function in equation 2.3 becomes:

$$f(\mathbf{R}) = \sum_\alpha \int_{R_\alpha}^{K_\alpha} K_\alpha - x_\alpha dx_\alpha = \sum_\alpha \frac{1}{2}(R_\alpha - K_\alpha)^2. \tag{2.15}$$

This is just ordinary quadratic programming. For the MacArthur's consumer resource model with linear resource dynamics (equation 1.18 with $\tau_\alpha = 1$), one has

$$f(\mathbf{R}) = \sum_\alpha \int_{R_\alpha}^{K_\alpha} \frac{K_\alpha - x_\alpha}{x_\alpha} dx_\alpha = \sum_\alpha [K_\alpha \log \frac{\mathrm{K}_\alpha}{\mathrm{R}_\alpha} + \mathrm{R}_\alpha - \mathrm{K}_\alpha]. \tag{2.16}$$

This functional form is just the Kullback–Leibler divergence, a commonly used similarity measure used in machine learning [Bis06].

With a convex objective function and known constraints, the steady states can be obtained through typical convex optimization packages, for instance, CVXPY in Python [AVDB18]. This results in significant improvement in numerical simulations since convex optimization algorithms converge to their minima between one and two orders of magnitude faster than direct numerical integration of the corresponding ODEs.

2.3 Extend for arbitrary niche models

The relation $q_{i\alpha} = -\frac{\partial g_i(\mathbf{R})}{\partial R_\alpha} R_\alpha$ cannot always be satisfied for choices of $q_{i\alpha}$. One common way this happens is if an organism affects the resource dynamics in ways that are unrelated to their own growth rate, whether by producing novel byproducts (cross-feeding), or by consuming resource types that do not limit their growth.

To solve this issue, in [MICM20], we show that the minimization principle can be extended to a much larger class of niche models that do not satisfy the stringent requirement that $q_{i\alpha} = -\frac{\partial g_i(\mathbf{R})}{\partial R_\alpha} R_\alpha$. To do so, we separate $q_{i\alpha}$ into a symmetric term and a remaining antisymmetric term:

$$q_{i\alpha} = q^S_{i\alpha} + q^A_{i\alpha}, \quad q^S_{i\alpha} = -R_\alpha \partial g_i / \partial R_\alpha. \tag{2.17}$$

As the expressions for $q_{i\alpha}$ and $q^S_{i\alpha}$ are known, we also know $q^A_{i\alpha} = q_{i\alpha} + R_\alpha \partial g_i / \partial R_\alpha$. Without loss of generality, substituting above equations into the general equation for the resource dynamics of equation 2.12, we obtain

$$\frac{dR_\alpha}{dt} = h_\alpha(\mathbf{R}) + \sum_i N_i q^A_{i\alpha}(\mathbf{R}) - \sum_i N_i \frac{\partial g_i}{\partial R_\alpha} R_\alpha. \tag{2.18}$$

We can rearrange the above equation into the form

$$\frac{dR_\alpha}{dt} = \bar{h}_\alpha(\mathbf{R}) - \sum_i N_i \frac{\partial g_i}{\partial R_\alpha}. \tag{2.19}$$

where we repalce $h_\alpha(\mathbf{R})$ by an effective resource dynamics

$$\bar{h}_\alpha(\mathbf{R}) = h_\alpha(\mathbf{R}) + \sum_i N_i q^A_{i\alpha}(\mathbf{R}) R_\alpha.$$

Note that there still exists unknown variables $\mathbf{N}$ and $\mathbf{R}$. However, we only care about the steady states $(\bar{\mathbf{N}}, \bar{\mathbf{R}})$ and thus replace the time-dependent variables with the steady state population abundance,

$$\bar{h}_\alpha(\mathbf{R}) = h_\alpha(\mathbf{R}) + \sum_i \bar{N}_i q^A_{i\alpha}(\bar{\mathbf{R}}). \tag{2.20}$$

If $\bar{\mathbf{N}}$ and $\bar{\mathbf{R}}$ are known, we can solve for the steady states with the new objective function,

$$\bar{f}(\mathbf{R}) = \sum_\alpha \int_{R_\alpha}^{K_\alpha} \left[h_\alpha(\mathbf{x}) + \sum_i \bar{N}_i q^A_{i\alpha}(\bar{\mathbf{R}}) \right] \frac{dx_\alpha}{x_\alpha}. \tag{2.21}$$

The reason this procedure can not be applied directly is that we do not *a priori* know the steady state values of the species or resources $(\bar{\mathbf{N}}, \bar{\mathbf{R}})$. If we knew the values in the first place, there would be no need to solve these equations at all. This problem of minimizing an objective function whose parameters depend on the solution arises frequently in Machine Learning, in the context of fitting models with latent variables [MBW+19, Bis06]. It can be solved with a simple iterative approach, called Expectation Maximization (EM), where one starts by guessing the values of these parameters, then minimizes the function, and then updates the estimates using the new solution. This procedure results in the Algorithm shown in Algorithm 1.

2.3.1 Application to microbial consumer resource model

We now show that this algorithm can be applied to solve for the steady-states of the microbial consumer resource models (MicroCRM) introduced in Section 1.3. The MicroCRM does not obey the relations in equation 2.13 because the production of metabolites breaks the symmetry of the effective interactions in the original consumer resource model. The dynamics of the MicroCRM (equation 1.19 with $\tau_\alpha - = 1$ for simplicity) are

Algorithm 1 Expectation Maximization (EM) Algorithm

Initialization: randomly initialize $\bar{\mathbf{N}}, \bar{\mathbf{R}}$, $\mathbf{N}, \mathbf{R}$
while $|\mathbf{R} - \bar{\mathbf{R}}| < \epsilon$ **do** ▷ ϵ controls the precision of the numerical solution.
$\bar{\mathbf{N}}, \bar{\mathbf{R}} \leftarrow \mathbf{N}, \mathbf{R}$
Expectation Step:
$\bar{f}(\mathbf{R}) = \sum_\alpha \int_{R_\alpha}^{K_\alpha} \left[h_\alpha(\mathbf{x}) + \sum_i \bar{N}_i q_{i\alpha}^A(\bar{\mathbf{R}})\right] \frac{dx_\alpha}{x_\alpha}$.
Maximization Step:
$\bar{\mathbf{R}} \leftarrow$ maximize $\bar{f}(\mathbf{R})$ subjected to constraints $g_i(\mathbf{R}) \leq 0, R_\alpha \geq 0$.
$\bar{\mathbf{N}} \leftarrow$ KKT multipliers corresponding to the constraints $g_i(\mathbf{R}) \leq 0$
end while
return $\bar{\mathbf{N}}, \bar{\mathbf{R}}$.

described by the equations

$$\begin{aligned} \frac{dN_i}{dt} &= N_i \left[\sum_\alpha (1 - l_\alpha) C_{i\alpha} R_\alpha - m_i\right], \\ \frac{dR_\alpha}{dt} &= \kappa_\alpha - R_\alpha - \sum_i N_i C_{i\alpha} R_\alpha + \sum_{\beta,i} D_{\alpha\beta} l_\beta N_i C_{i\beta} R_\beta. \end{aligned} \tag{2.22}$$

Substituting these equations into equations 2.13 and 2.17 yields:

$$q_{i\alpha}^S = -(1 - l_\alpha) C_{i\alpha} R_\alpha, \quad q_{i\alpha}^A = \sum_\beta D_{\alpha\beta} l_\beta N_i C_{i\beta} R_\beta - l_\alpha C_{i\alpha} R_\alpha \tag{2.23}$$

Assuming the steady state $\bar{\mathbf{N}}$ and $\bar{\mathbf{R}}$ are known, the effective resource dynamics is

$$\bar{h}_\alpha(\mathbf{R}) = \kappa_\alpha - R_\alpha + \sum_{i,\beta} \bar{N}_i (D_{\alpha\beta} l_\beta \bar{N}_i C_{i\beta} \bar{R}_\beta - l_\alpha C_{i\alpha} \bar{R}_\alpha) \tag{2.24}$$

As the second term is a constant, it is equivalent to replace κ with

$$\bar{\kappa}_\alpha = \kappa_\alpha + \sum_{i,\beta} \bar{N}_i (D_{\alpha\beta} l_\beta \bar{N}_i C_{i\beta} \bar{R}_\beta - l_\alpha C_{i\alpha} \bar{R}_\alpha). \tag{2.25}$$

The objective function still keeps the form as equation 2.16 but replace the upper integral limit with $\bar{\kappa}_\alpha$. Then, the microbial consumer resource model can be solved with the EM Algorithm 1.

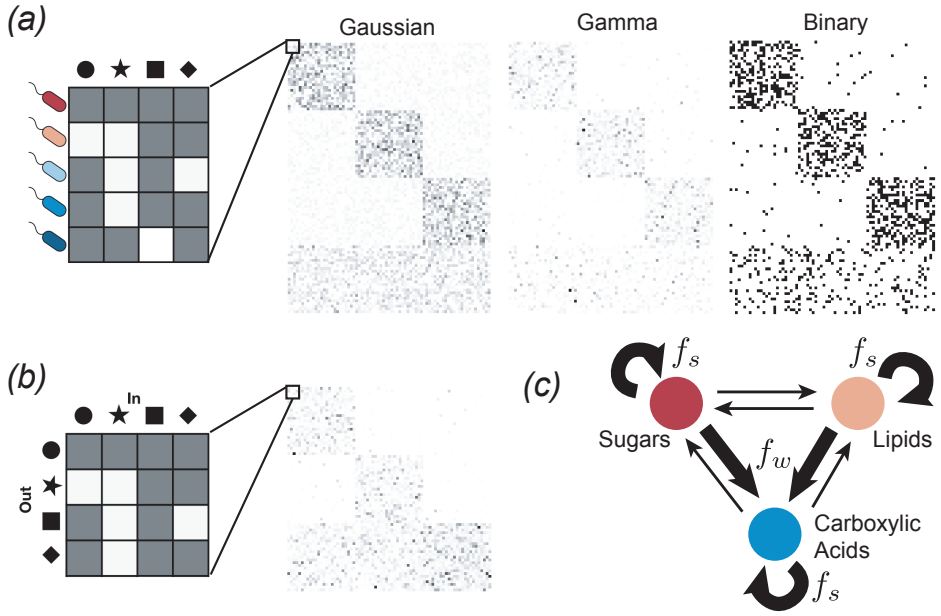

FIGURE 2.2: Sampling parameters and adding metabolic structure. *(a)* Sampling the consumer matrix $C_{i\alpha}$. An example of each of the three sampling choices is shown, with white pixels representing $C_{i\alpha} = 0$ and darker pixes representing larger values. The examples have $F = 3$ consumer families with specialism level $q = 0.9$, each with $S_A = 25$ species, plus a generalist family with $S_{\text{gen}} = 25$ species. *(b)* Sampling the metabolic matrix $D_{\alpha\beta}$. Each column represents the allocation of output fluxes resulting from metabolism of a given input resource. This example has $T = 3$ resource classes, and an effective sparsity $s = 0.05$. *(c)* Diagram of three-tiered metabolic structure. A fraction f_s of the output flux is allocated to resources from the same resource class as the input, while a fraction f_w is allocated to the "waste" class (e.g., carboxylic acids). In the example of the previous panel, allocation fractions were $f_s = f_w = 0.49$. Made by Robert Marsland III in [MCGM20]

2.4 Comparison with experimental observations

A major goal in ecology is to identify general principles shaping microbial ecosystems. In order to avoid unnecessary time-consuming and expensive wet lab experiments, one promising approach is to use minimal mathematical models to reproduce and understand experimentally observed ecological patterns [MCM20]. Here, we show that the Microbial Consumer Resource Model (MiCRM) (see Section 1.3) can reproduce patterns found in large-scale survey data, including the Earth Microbiome Project (EMP) and the Human Microbiome Project (HMP). Our model can help explain mechanisms resulting in patterns observed at the species scale.

2.4.1 Model assumptions

MiCRM considers the exchange and consumption of metabolites by introducing consumer matrix $C_{i\alpha}$ and metabolic matrix $D_{\alpha\beta}$. Typically, the elements in these matrices can be determined by measuring the specie's growth rate in different pair of two-species or species-resource coculture in experiments. However, a diverse community consists of hundreds of species and resources, suggesting to measure several thousands of parameters, which is unfeasible. For simplification, we just assume all parameters are sampled from a random distribution, for instance, the consumer matrix $C_{i\alpha}$, the maintenance cost m_i are sampled from Gaussian distribution. We encode the energy conservation in sampling the metabolic matrix $D_{\alpha\beta}$ by using a Dirichlet distribution, whose elements in the column are summed to be 1. κ_α is a nonzero value only for $\alpha = 0$ as we expect few resources are available in the environment.

Actually, $C_{i\alpha}$ and $D_{\alpha\beta}$ are not purely random as we have to incorporate metabolic and taxonomic structure at different levels by assuming species or resources from the same family have similar consumer preferences or byproduct stoichiometry. This is a reasonable biological assumption , for example, it is well known that bacteria from the Enterobacteria family have a strong preference for fermenting sugars. To capture this, we assign the M resources to T classes (e.g. sugars, amino acids, etc.), each with M_A resources where $A = 1, \dots T$ and $\sum_A M_A = M$. Likewise the total S_{tot} species can be assigned to F families, with $F \leq T$, and each family preferentially consuming resources from a different resource class. A generalist family can also be included, with S_{gen} species and no preferred resource class, so that $S_{\text{gen}} + \sum_A S_A = S_{\text{tot}}$. These setups result in the block structures in $C_{i\alpha}$ and $D_{\alpha\beta}$ in Figure 2.2.

In practice, we choose the metabolic matrix $D_{\alpha\beta}$ according to a three-tiered secretion model illustrated in Figure 2.2 (c). The first tier is a preferred class of 'waste' products, such as carboyxlic acids for fermentative and respiro-fermentative bacteria, with M_w members. The second tier contains byproducts of the same class as the input resource. For example, this could be attributed to the partial oxidation of sugars into sugar alcohols, or the antiporter behavior of various amino acid transporters. The third tier includes everything else. We encode this structure in $D_{\alpha\beta}$ by sampling each column β of the matrix from a Dirichlet distribution with concentration parameters $d_{\alpha\beta}$ that depend on the byproduct tier, so that on average a fraction f_w of the secreted flux goes to the first tier, while a fraction f_s goes to the second tier, and the rest goes to the third. The Dirichlet distribution has the property that each sampled vector sums to 1, making it a natural way of randomly allocating a fixed total quantity (such as the total secretion flux from a given input). To write the expressions for these parameters explicitly, we let $A(\alpha)$ represent the class containing resource α, and let w represent the 'waste' class. We

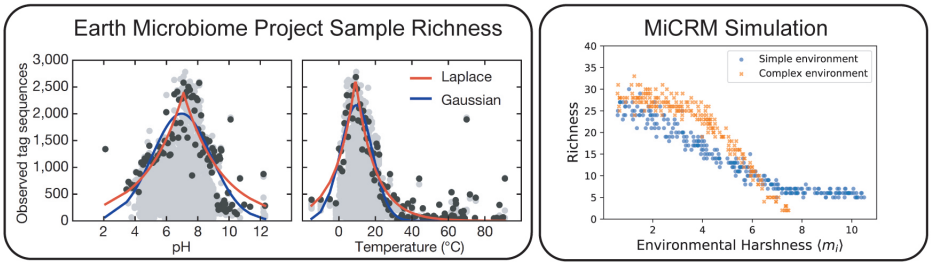

FIGURE 2.3: Relationship between diversity and environmental harshness is modulated by environmental complexity. Left: Gray dots are the number of distinguishable strains observed in each sample of the EMP, plotted vs. pH and temperature. Black dots represent the 99th percentile of all communities at a given pH or temperature. Colored lines are fits of a Laplacian and a Gaussian distribution to the 99the percentile points. Reproduced from Figure 2 of the initial open-access report on the results of the EMP[TSM+17]. Right: The number of species surviving to steady state in simulated communities, plotted vs. environmental harshness. Harsher environments at extreme pH or temperature were simulated by increasing the total amount of resource consumption m_i required for growth (by the same amount for all species). Blue squares are simulation results when all the energy was supplied via a single resource type, while orange circles are simulations where the incoming energy was evenly divided over all 90 possible resource types. Made by Robert Marsland III in [MCM20]

also introduce a parameter s that controls the sparsity of the reaction network, ranging from a dense network with all-to-all connection when $s \to 0$, to maximal sparsity with each input resource having just one randomly chosen output resource as $s \to 1$. With this notation, we have

$$D_{\alpha\beta} = \mathrm{Dir}(d_{1\beta}, d_{2\beta}, d_{3\beta}, \ldots, d_{M\beta})_\alpha \tag{2.26}$$

$$d_{\alpha\beta} = \begin{cases} \frac{f_w}{sM_w}, & \text{if } A(\beta) \neq w \text{ and } A(\alpha) = w \\ \frac{f_s}{sM_{A(\beta)}}, & \text{if } A(\beta) \neq w \text{ and } A(\alpha) = A(\beta) \\ \frac{1-f_s-f_w}{s(M-M_{A(\beta)}-M_w)}, & \text{if } A(\beta) \neq w \text{ and } A(\alpha) \neq A(\beta) \\ \frac{f_w+f_s}{sM_w}, & \text{if } A(\beta) = w \text{ and } A(\alpha) = w \\ \frac{1-f_w-f_s}{s(M-M_w)}, & \text{if } A(\beta) = w \text{ and } A(\alpha) \neq w. \end{cases} \tag{2.27}$$

The final two lines handle the case when the 'waste' type is being consumed. For these columns, the first and second tiers are identical. This led to an ambiguity in the expression presented in the Supporting Information of [MICG+19], which we have now clarified by treating this case separately. Note that in the third line, it is implicit that $A(\alpha) \neq w$, since $A(\alpha) = w$ is covered in the first line. For more simulation and model details, we refer interested readers to [MCGM20].

2.4.2 Patterns in the Earth Microbiome Project

The Earth Microbiome Project consists of over 20,000 samples in 17 different environments located on all 7 continents [TSM+17]. One of the interesting patterns is anti-correlation between richness and environmental harshness reproduced in Figure 2.3. Samples near neutral pH or at moderate temperatures ($\sim 15°C$) showed much higher richness than samples from more extreme conditions. Peak richness dropped by a factor of 2 for pHs less than 5 or greater than 9, and temperatures less than 5 °C or greater than 20 °C. The EMP samples also showed a strongly nested structure, namely, species appearing in lowly diverse communities tended to belong to one of the highly diverse communities. The nested structure is clearly shown in Figure 2.4, where each column in the matrix corresponds to a different sample and each row to a specific taxon.

These patterns may result from that microbes require higher maintenance cost to survive in harsher environments [HJ13]. For example, powering chaperones to prevent protein denaturation and running ion pumps to maintain pH homeostasis both require significant amounts of ATP. We hypothesized that varying m_i could explain the patterns observed in the EMP. In the MiCRM, m_i is sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. In order to model the varying harshness of the environment, we can sample the mean value by sampling the mean value m uniformly between 0.5 and 10.5 with the same standard deviation. As a large m corresponds to harsh environments with increased energetic demands. the anti-correlation between richness and environmental harshness can be expected, shown in Figure 2.3.

Surprisingly, the same simulation also captures the nestedness of the EMP data, shown in Figure 2.4. To confirm the nested pattern results from harshness variations, we ran simulations with the varying dispersal limitation, i.e., the initial number of species from the regional species pool allowed to colonize the community was randomly chosen. In the new simulations, shown in the bottom right panel of Figure 2.4, the nestedness vanishes, suggesting nestedness may be a sign of selection-dominated community assembly.

2.4.3 Patterns in the Human Microbiome Project

The Human Microbiome Project is a large-scale survey of the microbial communities that reside in and on the human body [HGK+12a]. Here we discuss two major patterns in the human microbiome, shown in the top half of Figures 2.5 and 2.6. First, for a given body site, different individuals had similar community composition patterns in the phylum level (see Fig. 2.5). But samples from different body sites typically differed more than samples from the same body site, leading to the second pattern, shown

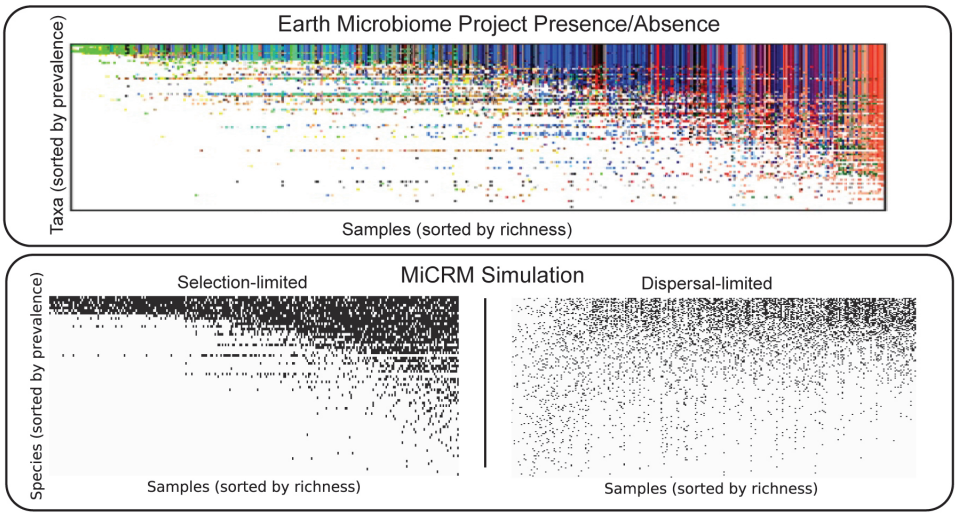

FIGURE 2.4: Nestedness of community composition indicates selection-dominated community assembly. Top: Presence (colored) or absence (white) of each microbial phylum in a representative set of 2,000 samples from the EMP. Reproduced from Figure 3 of the EMP report [TSM+17]. Different colors represent different biomes. Bottom: Presence (black) or absence (white) of species in simulated communities. Two different regimes of community assembly were simulated. The first is the selection-dominated scenario of Figure 2.3, where variability in diversity is produced by variations in environmental harshness, and all samples are initialized with the vast majority (150/180) of the species in the regional pool. The second is a dispersal-dominated scenario, where environmental conditions are identical for all samples, but each sample is initialized with a different number of species, varying from 1 to 180. See main text and Methods for simulation details. Made by Robert Marsland III in [MCM20]

in Figure 2.6 of clustering of microbial communities by body site across individuals [HGK+12a, QLR+10], which suggests the sample location can be inferred by the relative abundance data.

One important factors is that different kinds of externally supplied nutrients, such as fibers and proteins, are thought to encourage growth of different microbial taxa. For this reason, we hypothesized that the patterns in the HMP may arise from heterogeneity in the resources available in different environments. In order to reproduce such patterns, it is important to assume some minimal level of taxonomic and metabolic structure. As a result, we divided resources into six resource classes and species into six families, with each family specializing in one resource class, as illustrated in Figure 2.2 and described above.

We first assumed there were only two externally supplied resources. In particular, the three different "body sites" was modeled by supplying with a unique pair of resources from distinct resource classes (i.e. body site 1 was supplied with a resource from class A and a resource from class B, body site 2 with a resource from class C and a resource from class D, and body site 3 with a resource from class E and a resource from class

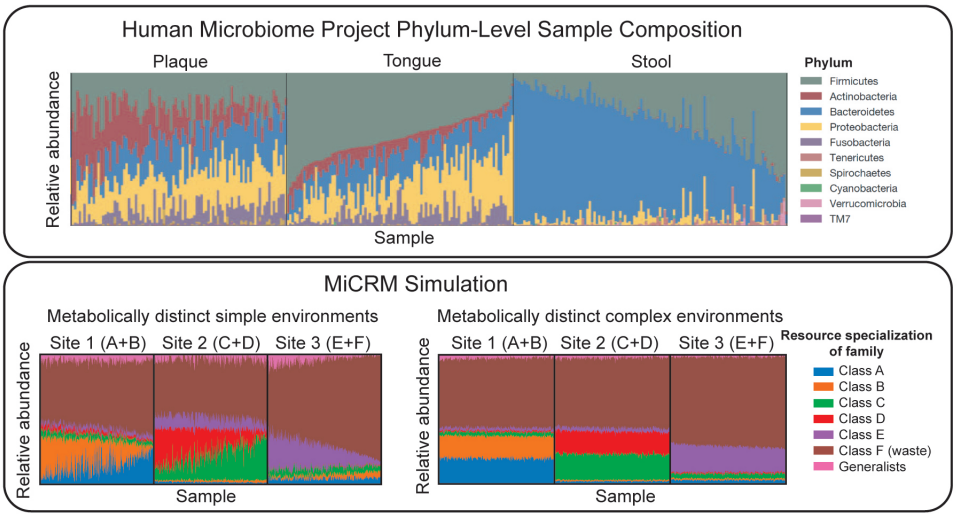

FIGURE 2.5: Low-dimensional nutrient supply variation reproduces patterns in human microbiome survey data. Top: Each column represents one sample from the Human Microbiome Project (HMP). Colored segments represent relative abundances of different phyla in each community. Reproduced from Figure 2 of the initial open-access report on the results of the HMP[HGK+12a]. Bottom: Each column represents one of 900 simulated samples, each stochastically colonized with 2,500 species from a regional pool of 5,000 species, comprising seven metabolically distinct families. Colored segments represent relative abundances of the seven families defined in Figure 2.2. Each of the three "body sites" was supplied with resources from a different pair of resource classes, with total nutrient supply fixed. In the first set of simulations (left), one resource from each class was supplied, and the ratio of the two supply rates was randomly varied from sample to sample. In the second set (right), all resources from each class were supplied, with randomly chosen supply rates for each sample, normalized to keep the total supply fixed. The brown family present in all three environments specializes in the typical byproducts (e.g., carboxylic acids) generated from all the other resource classes. Within each body site, samples are sorted by relative abundance of this family. See main text and Methods for simulation details. Made by Robert Marsland III in [MCM20]

eps2

F). We modeled variability in the availability of resources across individuals at a fixed body site by changing the ratio of the two supplied resources while holding the total supplied energy fixed. We generated a regional pool of 5,000 species (approximately the number of OTU's identified in the HMP [HGK+12a]), and stochastically colonized 300 samples per body site with 2,500 species each. Figure 2.5 shows the resulting patterns for simple (two externally supplied resources from different classes) and complex environments (supplied with 100 randomly chosen distinct resources regardless of resource class). For simple environments, our simulations reproduced the patterns exhibited in the data including gradients in the dominant families present at each of the body sites. In contrast, for complex environments we see that the relative abundance of different families stays almost constant across individuals for each body site. This suggests that the patterns found in Fig 2.5 may reflect the combined effects of environmental filtering

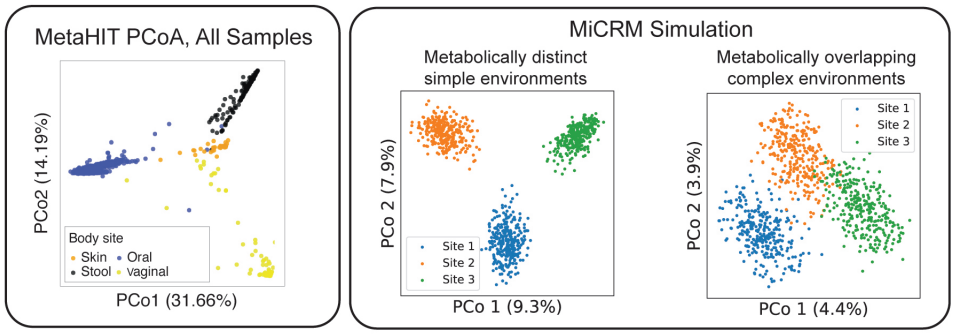

FIGURE 2.6: **Correlations between inter-site nutrient variation and metabolic structure affect distinguishability of body sites.** Left: Principal coordinate analysis (PCoA) of MetaHIT OTU-level community compositions, using the Jensen-Shannon distance metric. Data points are colored by the body site from which the sample was taken. Reproduced from Figure 1 of [CHM+18]. Right: Jensen-Shannon PCoA of species-level compositions of the simulated communities. In the first set of simulations (left), the nutrients supplied to different body sites come from different resource classes. In the second set of simulations (right), each environment is supplied with a randomly chosen set of resource types, with each site being supplied with about one third of the 300 possible resources. Made by Robert Marsland III in [MCM20].

and competition between species in the presence of a few dominant externally supplied resources.

We can also perform a PCoA across body sites to the data from simulations, as in the MetaHIT data. As can be seen in Figure 2.6, these simulations recapitulated the pattern seen in real microbial communities. We found that this clustering by body site depended strongly on the fact that different body sites had metabolically distinct resources. For example, the clusters were no longer fully separable on a two-dimensional PCoA for a complex environment (right most graph in Figure 2.6). This suggests that the clustering of human microbiomes according to body-sites likely reflects the fact that these body sites have metabolically distinct environments that result in different patterns of byproduct secretion.

In summary, our analysis suggests several hypotheses relating mechanism to large scale patterns observed in both the EMP and HMP. We show it is possible to reproduce these patterns with a minimal mathematical model and quantitatively understand patterns at both the species and community levels.

Chapter 3

Statistical-physics-inspired approaches for complex ecosystems

Nature has revealed an astounding degree of phylogenetic and physiological diversity in microbial communities. Recent advances in DNA sequencing technologies makes it possible to measure microbial communities abundances at high resolution, opening a new precision era in microbial ecology [CLW+11, SWGV14]. Understanding such large amounts of microbial data challenges current theories and analytical approaches, most of which were developed using models that consider only a few species. This challenge motivated us to develop new theoretical approaches for understanding ecology directly in a high-dimensional setting by analyzing ecological models with a large number of species and resources.

In Chapter 1, our discussion and examples were limited to ecosystems consisting of few species. In low dimensions, describing every degree of freedom (e.g. resources and species abundances) is tractable. Community properties can be evaluated by exhaustively searching all states directly. However, most microbial ecology dataset are high dimensional, simultaneously measuring the relative abundances of hundreds of species across different habitats. A huge number of combinations of states must be considered, and the use of exhaustive search strategies is no longer feasible, i.e., one suffers from *the curse of dimensionality* [Ric57].

One pertinent example in statistical mechanics is the simple example of *particles in a box*. At any time t, the microstate of system of N particles is described by the positions $\vec{x}_i(t)$ and the momenta $\vec{p}_i(t)$ of the different particles. When N is not too large, the

evolution of a microstate can be predicted precisely by solving Hamilton's equations for all particles. When N becomes large, for example, a typical volume of gas has the order of $N_A \sim 10^{23}$ particles, the complexity of predicting the microstate is too high to be feasible.

Taking the analogy between particles in a box and species in the ecosystem, analytical approaches and theoretical insights derived from small ecosystems may not scale up to large ecosystems. Instead, one should look for theoretical frameworks that directly work in the high-dimensional setting. This suggests that we should develop a statistical mechanics approach to complex ecosystems.

3.1 Large N limit and typicality

We have stressed the practical difficulties in high dimension. Actually, taking the thermodynamic limit $N \to \infty$ can also lead to a number of simplifications. As we all know, statistical mechanics has successfully dealt with the system with a very large number of degrees of freedom [Ma18]. For the *particles in a box* example, instead of trying to predict the behavior of individual particles in a box, we can make quite accurate statements about the macro behavior averaged over millions of particles. Because typical features of the gas can emerge from the collective behavior of many individuals, we can often even ignore many details about the identity of the particles themselves. For example, the macroscopic quantities, such as pressure, temperature, volume and entropy follow Maxwell relations, no matter if the gas is formed by oxygen, nitrogen or mixed. In order to derive typical relations in thermodynamics, statistical physicists have to assume some general principles. For example, in order to obtain the Boltzmann distribution, we have to assume the maximum entropy principle, which means thermal systems are at the largest uncertainty, i.e. every microstate is identical and has the same probability.

Taking the analogy between species in ecosystems and particles in box, we can learned three lessons from statistical mechanics that we can apply to ecological systems:

1. We need to find typical relations between macro properties of the ecosystem, such as the relation between statistical properties of the species and resources, rather than functions and behaviors of one particular species.

2. Some principle must be assumed in order to obtain these typical relations. For example, the ecosystem is disordered (meaning parameters can be modeled as being drawn from a random distribution). With this simple but important assumption, we can apply physics-inspired approaches to derive many interesting relations in complex ecosystems.

3. The typical relations depends on the ecological dynamics. We will solve Lotka-Volterra and consumer-resource models with cavity method, and then can show how distinct macro behaviors emerge depending on which dynamics we use.

3.2 Disorder ecosystems

Why can we treat ecosystems as disorder systems? Starting with Robert May's pioneering work in 1972 [May72], there is a long tradition in theoretical ecology of treating ecosystems like disorder systems. In order to apply the physics-inspired approaches introduced in this chapter, we have to assume the species-species or species-resource interactions are random. In reality, species in the same family share similar metabolic pathways, leading to correlations among parameters [GLB+18]. However, we still make this simplifying assumption for three reasons. First, it greatly simplifies the mathematics. While more realistic taxonomic structures with correlated parameters can be analyzed with advanced mathematical tools, such as replica symmetry breaking in spin glass theory [MPV87], this is quite mathematically challenging. Second, we can learn a lot from a simple models. In the history of solid state physics, people worked on understanding Ising model, which only considers the nearest neighbor interactions and assumes all interactions have the same amplitude. In real materials, the interactions are more complex and these assumptions do not hold. Third, our work shows that complex ecosystem may actually behave as if they were random even if they are structured because introducing even a small amount of noise can cause a phase transition to "typicality" [CMIM19]. Given a deterministic structure of the consumer preference matrix (an identity or a block matrix), we show that adding even a small amount of noise in the consumer preferences (proportional to the inverse system size, $1/S$) will destroy the engineered structure and make the macroscopic properties of an ecosystem indistinguishable from a random ecosystem. This suggests complex ecosystems can be treated like disordered systems as long as we are concerned with predicting macroscopic ecological properties that reflect averages over many species and/or resources. We provide a mathematical proof of these statements in Chapter 5.

3.3 Random matrix theory

The eigenvalues of the interaction matrix play an important role in the dynamics of an ecosystem. Random matrix theory(RMT) can tell us what the eigenvalue spectrum look like analytically, when the matrix size is large and all its elements are independent and identically sampled from a distribution whose tail is exponentially bounded [Tao12]. For

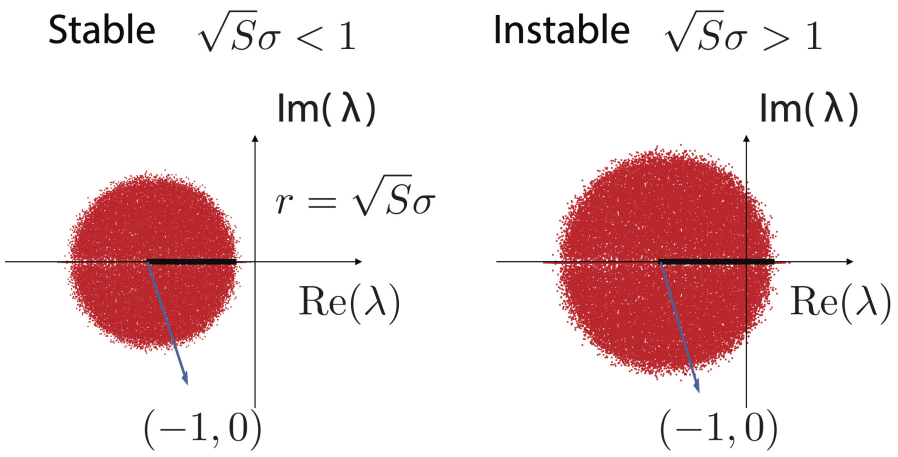

FIGURE 3.1: Schematic for May's stability criteria. The red scatter points are the eigenvalues on the complex plane.

a Gaussian random matrix $\mathbf{A}$, whose elements are sampled independent and identically sampled from a Gaussian distribution, the spectrum follows Girko's Circular Law and all its eigenvalues distribute uniformly on a circle in the complex plane [Gin65]. If $\mathbf{A}$ is a symmetric Gaussian random matrix, it follows Wigner Semicircle Law and all its eigenvalues distribute uniformly on a semicircle in the real plane [Wig58]. If it is a random consumer matrix $\mathbf{C}$, which is not square, RMT tells its covariance matrix $\mathbf{C}\mathbf{C}^T$ follows Marchenko–Pastur distribution [MP67a].

An amazing observation is that the eigenvalues of a RMT are distributed over a bounded domain. The upper and lower bounds are determined by the variables in the corresponding sampled probability distribution. As we know, the largest/smallest eigenvalue determines the stability of a fixed point in a dynamical system. This suggests there exists a stability-instability phase transition, and the critical point defining the transition can be predicted by RMT. In Section 3.3.1, introduce the May's famous stability criteria [May72] and describe its relation to RMT. In Chapter 5, we use recent progress in RMT to explain why complex ecosystems tend to behave as if they were completely disordered.

3.3.1 May's Stability Criteria

For an ecosystem of S species, May's theorem concerns the $S \times S$ community matrix $\mathbf{J}$, whose entries J_{ij} describe how much the growth rate of species i is affected by a small change in the population N_j of species j from its equilibrium value. The population

dynamics is obtained by Taylor expansion around some equilibrium point $\bar{N}_j$,

$$\frac{dN_i}{dt} = \sum_j J_{ij}(N_j - \bar{N}_j) \tag{3.1}$$

From above equation, the stability of this equilibrium can be quantified in terms of the largest eigenvalue $\lambda_{\max}$ of $\mathbf{J}$. If $\lambda_{\max}$ is positive, the equilibrium is unstable, and a small perturbation will cause the system to flow away from the equilibrium state.

In the 1960's, Jean Ginibre derived a mathematical formula for the distribution of eigenvalues in a special class of large random matrices[Gin65]. Girko's Circular Law states when the J_{ij} are sampled independently from probability distributions with zero mean and variance σ^2, taking the limit $S \to \infty$, its eigenvalues are uniformly distributed on a disk with radius $r = \sqrt{S}\sigma$ in the complex plane, shown in Fig. 3.1. And thus, the largest eigenvalue $\lambda_{\max}$ is at the boundary of this disk. With this result in mind, May considered a simple ecosystem where each species inhibits itself, with $J_{ii} = -1$, but different species initially do not interact with each other. This ecosystem is guaranteed to be stable for any level of diversity. He then examined how the stability is affected by adding randomly sampled interactions, and found that $\lambda_{\max}$ typically becomes positive when the root-mean-squared total strength $\sqrt{S\sigma^2}$ of inter-specific interactions reaches parity with the intra-specific interactions, which gives May's stability criterion:

$$\sqrt{S\sigma^2} = 1. \tag{3.2}$$

For a given pairwise interaction strength σ, this relation gives that the diversity S promotes ecosystem instability and large ecosystems tend to be unstable.

Before the 1970s, ecologists believed that diversity enhanced ecosystem stability. May's stability criteria challenged this idea and led to what is now commonly know as the diversity-stability debate [McC00]. Theorists have tried to circumnavigate May's original argument by changing the May's admittedly non-realistic assumptions, including by adding biologically realistic correlation structures [AT12], modular structures [GRA16], and correlations [AT15] to the interaction matrix, incorporating the dependence of the community matrix on population sizes [GGRA18], and considering high-order interactions in Lotka–Volterra dynamics [GBMSA17].

3.4 Spin-glass-inspired approaches

Recently, it is recognized that the generalized Lotka-Volterra model can be approximately reduced to a presence-absence model by replacing the absolute value of species

abundance N_i with a binary variable S_i, where $S_i = 1$ if the species is present and $S_i = 0$ if it is absent [FM14, DFM16]. By making an analogy between the presence/absence of a species in an ecosystem and whether a spin is up or down, it is possible to map ecological dynamics to the dynamics of spin models in statistical physics, which have been studied for more than one century [Isi25]. For disorder ecosystems, understanding community coexistence pattern becomes a spin glass problem and motivates physicists to use insights from spin glass theory to uncover the universal features of complex ecosystems. The spin-glass-inspired approaches are divided into two categories: cavity method (see Chapter 4) and replica method[MPV87, MM09]. These two methods have been proven to be equivalent to each other but formalize the same problem from very different perspectives [MP03]. Finally, we note that both the cavity method and replica methods have deep connections with RMT. We refer the interested readers to [RCKT08, LNV18, CRM20].

3.4.1 Replica Method

Before proceeding, for completeness we briefly discuss the Replica Method in the context of ecological dynamics. The replica method is another theoretical method originally developed to study spin glasses. The basic idea of replica method is to use replica trick $\log Z = \lim_{n\to 0} \frac{Z^n}{n}$ to estimates the partition function Z, which occurs in many problems of statistical mechanics and describes the statistical properties of a thermal-equilibrium system. The application of Replica method requires a predefined energy function. In Chapter 2, we have shown there is duality between a wide class of ecological dynamics and constrained optimization over Lyapunov functions. As constraints can also be expressed as energy functions, combining the Lyapunov function and constraints, the partition function can be estimated by replica method, and thus properties of ecological systems at steady states can be analyzed.

In principle, replica method is equivalent to the cavity method [MP03]. [BBC18b] shows replica approach can reproduce the cavity solution for Lotka-Volterra model in the unique equilibria phase and find the phase transition point from unique to multiple equilibria phase. These results were originally derived using cavity methods in [Bun17]. However, we note that a naive application of the cavity method fails in multiple equilibria phase and it becomes tedious to consider multiple-equilibria corrections. While replica method still works with one step replica-symmetry-breaking (RSB) approximation.

Another important application of replica method is to study feasibility in ecology. Structural stability gives a quantitative measurement of feasibility. It is challenging to estimate structural stability in high dimension because of the undersampling problem. Assuming ecosystems are disorder, replica method can analytically estimate structural stability since this quantity is closely related to properties of partition functions [GAS+17]. The formalism of underlying feasibility problems actually belongs to a class of constraint satisfaction problem, including random K-SAT and jamming problems [MPZ02, FP16], and this analogy has been discussed in [TM17, LE20]. In this thesis, we do not make use of the replica method, instead drawing on ideas and techniques from the cavity method.

Chapter 4

Cavity method for ecological models

The cavity method is not just limited to spin glasses. It has been successfully used for solving computer science problems, including K-SAT [MPZ02], compressed sensing [KMS+12] and combinatorial optimization [ZK16]. The original cavity method is designed for a model with binary variables. Over the last five or six years, tremendous progress has been made to develop and generalize this method to solve ecological models with non-negative continuous variables (e.g. species and resource abundance) [Bun17, ABM18a, BA17].

The basic idea behind the cavity method is to derive self-consistency equations by relating an ecosystem with S species / M resources to another ecosystem with S+1 species / M+1 resources. In the thermodynamic limit $S, M \to \infty$ there is no difference between observables computed in both systems. In other words, the statistical properties, such as the first and second moment of species abundance $\langle N \rangle$, $\langle N^2 \rangle$ and resource abundance $\langle R \rangle$, $\langle R^2 \rangle$ are the same for both systems. Adding a new "cavity" species/resource (which by convention we designate with the index 0) to the original ecosystem results in a perturbation to the original equilibrium state. In the cavity methods, one assumes that the adding a new "cavity" species/resource is small perturbation (of order $1/S$ or $1/M$) and hence the system before and after addition can be related using perturbation theory. Combining this with the idea of self-averaging allows for the derivation of a coupled set of self-consistent mean-field equations for the moments of the species and resource distributions.

The cavity method can predict coexistence patterns (species/resource abundances), stability, and species packing bounds (how many species can survive in an ecosystem) of

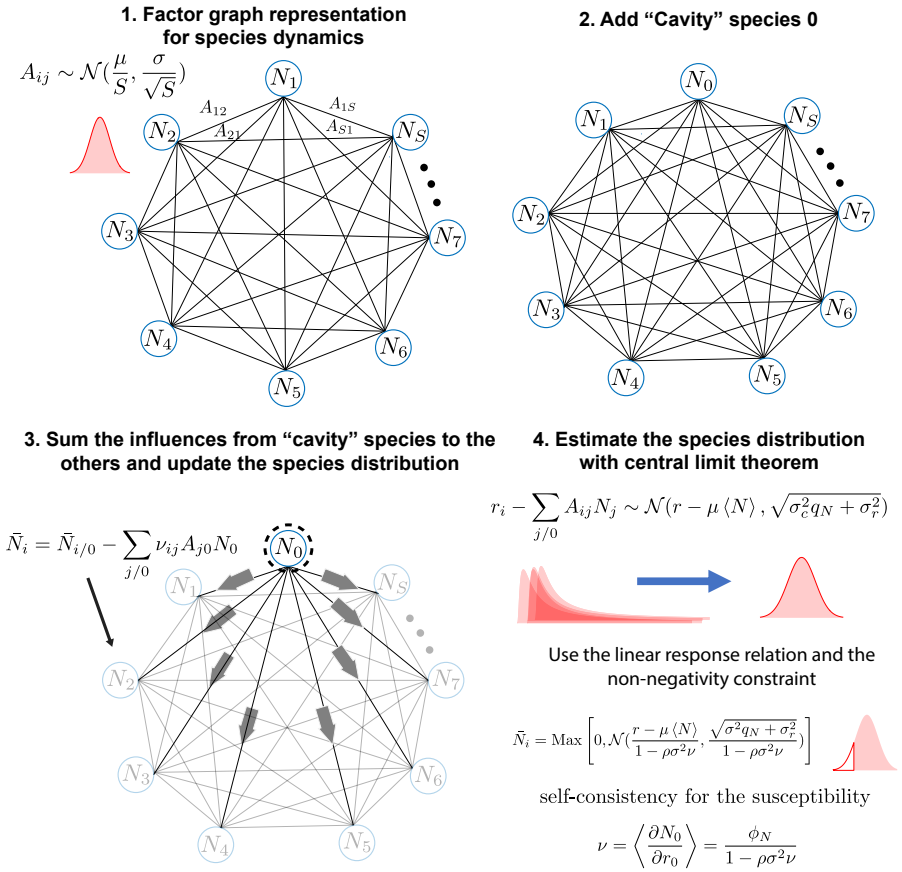

FIGURE 4.1: Schematic outlining steps in cavity solution for Lotka–Volterra model. **1.** The species dynamics in eq. (4.2) are expressed as a factor graph. The edges are bi-directional and sampled from a Gaussian distribution. **2.** Add the "Cavity" species 0 as the perturbation. **3.** Sum the resource abundance perturbations from the "Cavity" species 0 at steady state and update the species abundance distribution to reflect the new steady state. **4.** Employing the central limit theorem and the non-negativity constraint, the species distribution is expressed as a truncated normal distribution. The susceptibility appearing in the species distribution is the self-consistency relation.

a random ecosystem. We derive cavity equations for generalized Lotka-Volterra model and consumer-resource model in Section 4.1 and 4.2.

4.1 Cavity Method for Lotka–Volterra model

The following calculations follow [Bun17] closely. We consider the following generalized Lotka-Volterra equations consisting of S species,

$$\frac{dN_i}{dt} = g_i N_i (r_i - N_i - \sum_j A_{ij} N_j), \quad i = 1, 2, \ldots, S. \tag{4.1}$$

Here g_i are the intrinsic growth rates, r_i are the carrying capacities and A_{ij} encode inter-species interactions, with positive and negative values representing competition and mutualism interactions, respectively. We care about the statistical properties of the fixed point , where $\frac{dN_i}{dt} = 0$. Notice the intrinsic growth rates do not affect the fixed point and hence we set them equal to one in what follows. The steady-state equations become

$$0 = \bar{N}_i(r_i - \bar{N}_i - \sum_{j \neq i} A_{ij}\bar{N}_j). \tag{4.2}$$

where $\bar{N}_i$ are the species abundance in equilibrium.

The self-consistency equations are derived in the thermodynamics limit $S \to \infty$. In this limit, we assume many species present in the community have no correlation with each other so that the carry capacities r_i are independent and identically (i.i.d.) sampled from some probability distribution. For the interaction coefficients A_{ij}, we consider a correlation coefficient $\rho = \text{corr}(A_{ij}, A_{ij})$ with $-1 \leq \rho \leq 1$. $\rho = 1, -1$ correspond to completely symmetric and asymmetric interaction coefficients, respectively.

We assume that the A_{ij} are sampled from a Gaussian distribution with mean $\langle A_{ij} \rangle = \frac{\mu}{S}$ and variance $var(A_{ij}) = \frac{\sigma^2}{S}$. We can rewrite

$$A_{ij} = \frac{\mu}{S} + \sigma a_{ij} \tag{4.3}$$

where $\langle a_{ij} \rangle = 0$, $\langle a_{ij}a_{kl} \rangle_{i \neq j, k \neq l} = \frac{1}{S}\delta_{ik}\delta jl + \frac{\rho}{S}\delta_{il}\delta jk$. The scaling in the mean and variance is to keep the interaction term $\sum_j A_{ij}\bar{N}_j$ in eq. (4.2) independent of the system size. r_i is also sampled from another gaussian distribution with mean r and variance σ_r, independent of A_{ij}. We rewrite it as

$$r_i = r + \delta r_i, \tag{4.4}$$

where $\langle \delta r_i \rangle = 0$, $\langle \delta r_i \delta r_j \rangle = \sigma_r^2 \delta ij$.

Now we perturbed the original ecosystem with a new "cavity" species N_0 with interactions A_{0j} and A_{j0},

$$N_0(r_0 - N_0 - \sum_{j=1}^{S} A_{0j}N_j) = 0. \tag{4.5}$$

We represent the original steady state with S species is $\bar{N}_{i/0}$ and the new steady state with $S+1$ species with $\bar{N}_i$. The susceptibility function to the perturbation is defined by $\nu_{ij} = \frac{\partial N_i}{\partial r_j}$. Adding the "cavity" species N_0 is equivalent to decreasing r_i by $\sum_j A_{j0}N_0$. Note that in the thermodynamic dynamics limit $S \to \infty$, A_{0j} and A_{j0} scale as $\frac{1}{\sqrt{S}}$, other

$\mathcal{O}(1/S)$ terms are ignored. The perturbation equation becomes:

$$\bar{N}_i \quad = \quad \bar{N}_{i/0} - \sum_{j=1}^{S} \nu_{ij} A_{j0} N_0 \tag{4.6}$$

With the help of eq. (4.3) and eq. (4.4), solving eq. (4.5) yields:

$$N_0 = \max\left(0, \frac{r_0 - \sum_{j=1}^{S} \sigma a_{0j} N_{i/0} - \mu \langle N \rangle}{1 - \sigma^2 \sum_{j,i=1}^{S} \nu_{ij} a_{0j} a_{i0}}\right) \tag{4.7}$$

where $\langle N \rangle = \frac{1}{S} \sum_j N_j$. max is a function to take the maximum value in the bracket. The sum term in the denominator can be written by noting that to leading order in S that

$$\sum_{j,i=1}^{S} \nu_{ij} a_{0j} a_{i0} = \frac{\rho}{S} \sum_{j} \nu_{jj} = \rho\nu. \tag{4.8}$$

From the definition of susceptibility ν_{ij}, the self-consistency relation yields

$$\nu = \left\langle \frac{\partial N_0}{\partial r_0} \right\rangle = \frac{\phi_N}{1 - \rho\sigma^2\nu}. \tag{4.9}$$

where $\phi_N = \frac{S^*}{S}$ is the fraction of nonzero species in the ecosystem. Note that the factor of ϕ_N results from the fact that if $N_0 = 0$, the corresponding derivative in the susceptibility average is also zero.

Assuming N_i are weak correlated with each other, using the central limit theorem, the sum $\sum_{j=1}^{S} \sigma a_{0j} N_{i/0}$ in the numerator can be approximated to a Gaussian distribution with mean 0 and variance $\sigma_c^2 \langle N^2 \rangle$, where $\langle N^2 \rangle = \frac{1}{S} \sum_{i=1}^{S} N_i^2$. We rewrite eq. (4.10),

$$N_0 = \max\left(0, \frac{r - \mu \langle N \rangle + \sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle} z}{1 - \rho\sigma^2\nu}\right) \tag{4.10}$$

which is a truncated Gaussian. Here z is an auxiliary Gaussian variable with mean 0 and unit variance and $\sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}$ results from the sum of two Gaussian variables. Now we have unknown truncated Gaussian moments ϕ_N, $\langle N \rangle$ and $\langle N^2 \rangle$, which can be determined by the species abundance distribution eq. (4.10).

The following notations are helpful. Let $y = \max\left(0, \frac{a}{b} + \frac{c}{b}z\right)$, with z being a Gaussian random variable with zero mean and unit variance. Then its j-th moment is given by

$$\begin{aligned} \langle y^j \rangle &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(\frac{c}{b}x + \frac{a}{b}\right)^j dx \\ &= \left(\frac{c}{b}\right)^j \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(x + \frac{a}{c}\right)^j dx \\ &= \left(\frac{c}{b}\right)^j w_j(\frac{a}{c}) \end{aligned} \tag{4.11}$$

here we define $w_j(\frac{a}{c}) = \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(x + \frac{a}{c}\right)^j dx$.

With the help of w_j, we have all self-consistency equations,

$$\nu = \frac{\phi_N}{1 - \rho\sigma^2\nu}, \quad \phi_N = w_0(\frac{r - \mu \langle N \rangle}{\sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}}), \tag{4.12}$$

$$\langle N \rangle = \frac{\sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}}{1 - \rho\sigma^2\nu} w_1(\frac{r - \mu \langle N \rangle}{\sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}}), \tag{4.13}$$

$$\langle N^2 \rangle = \frac{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}{(1 - \rho\sigma^2\nu)^2} w_2(\frac{r - \mu \langle N \rangle}{\sqrt{\sigma_r^2 + \sigma^2 \langle N^2 \rangle}}). \tag{4.14}$$

Above self-consistency equations can be solved numerically in Mathematica. Fig. 4.2 show the Comparison between the cavity solution and 500 independent numerical simulations for various ecosystem properties such as the fraction of surviving species $\frac{S^*}{S}$ and the first and second moment of the species distribution for a symmetric ($\rho = 1$) and uncorrelated ($\rho = 0$) A_{ij} with $S = 200$, $\mu = 2.$, $r = 1.$, $\sigma_r = 0.1$. As can be seen in the figures, our analytic expressions agree remarkably well over a large range of σ_c.

4.1.1 Connection with May's Stability Criteria

Fig. 4.2 shows, in the symmetric case, cavity solution starts deviating from the numerical simulations when $\sigma_c > 0.5$. This is actually indicative of the emergence of a new phase where the replica symmetric ansatz used in our cavity calculations no longer holds. To understand this phase we can look at the minimum eigenvalue of the interaction matrix A_{ij}^* restricted to species that survive in the ecosystem at steady-state. Fig. 4.2 (D) shows the minimum eigenvalue of A_{ij}^* as a function of noise in the consumer preferences σ_c. The minimum eigenvalue decrease monotonically with increasing σ_c until it reaches zero and then the cavity solution fails. In the numerics, this happen slightly earlier than zero due to finite size effects. This is reminiscent of the scenario described by May's stability criteria discussed in Section 3.3.1. In May's case, there are only two phases: unique fixed point (UFP) and unbounded growth (UG), characterizing by λ_{min} larger or

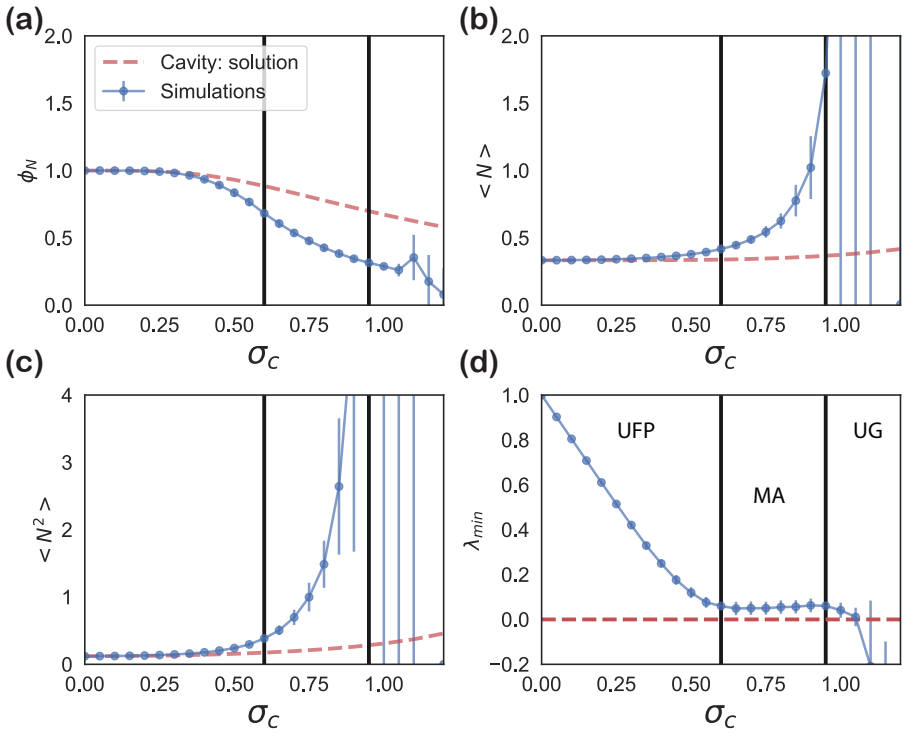

FIGURE 4.2: Comparison between the cavity solution (equation 4.12 - 4.14) and simulations for the fraction of surviving species $\phi_N = \frac{S^*}{S}$, (a) the fraction of surviving species $\phi_R = \frac{M^*}{M}$ and (b) the first moment $\langle N \rangle$ and (c) the second moment $\langle N^2 \rangle$ of the species distributions as a function of σ_c. (d) The minimum eigenvalue of the submatrix A^*_{ij} at at different σ_c. The error bar shows the standard deviation from 500 numerical simulations with $S = 200$, $\mu = 2.$, $r = 1.$, $\sigma_r = 0.1$ and $\rho = 1$.The black solid lines separate the results in three different regimes: unique fixed point, multiple attractors and unbound growth.

smaller than zero, as shown in Fig. 3.1. While in the full Lotka–Volterra dynamics, Fig. 4.2 (D) shows that there exists an additional multiple-attractors(MA) phase separating these regimes, where λ_{min} gets pinned to zero over a finite range of σ_c (see Fig. 4.3). This MA phase was first discovered in [Bun17].

In the MA phase, the dynamics system is marginally stable and highly sensitive to initial conditions. It has deep connections with the de Almeida-Thouless line in spin glass theories [dAT78] and chaotic behavior in random one-layer neural networks [SCS88]. The UFP-MA phase only happens in certain parameters regime, for example $\mu > 0$ and $\rho = 1$. In other regimes, the system may have a UFP-UG phase transition instead of a UFP-MA phase transition. The replica symmetric cavity approach can only predict the σ^* when $\lambda_{min} = 0$, but has no idea about whether λ_{min} will stay zero (MA) or become negative(UG) when further increasing σ_c. This is because the replica symmetry is broken in the MA and UG phases. Analysis beyond the UFP phase has been carried out using the replica approach with replica symmetry breaking [BBC18a].

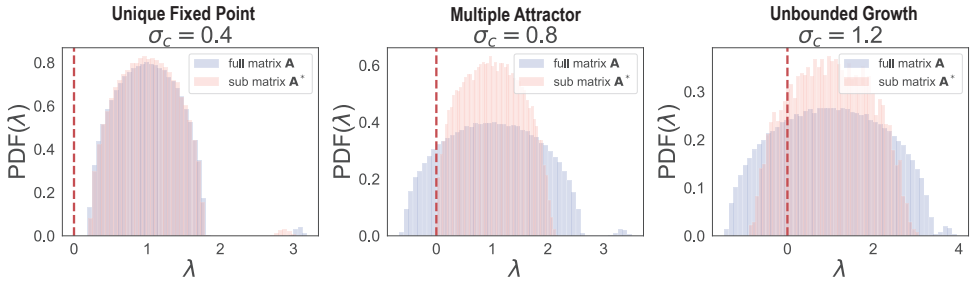

FIGURE 4.3: Spectrum of the whole species interaction matrix $\mathbf{A}$ and the surviving species interactions matrix $\mathbf{A}^*$ at the unique fixed point, multiple attractor and unbounded growth phase. The parameters are the same as Fig. 4.2.

What causes the difference in behaviors between LV dynamics eq. (4.1) and May's dynamics governed by eq. (3.1)? In May's case, we asked about the stability of the fixed point where all species survive for the special case of symmetric $\mathbf{A}$ and derived May's stability criteria:

$$4\sigma^2 = 1. \tag{4.15}$$

Note that this Eq. looks superficially different from eq. (3.2). However, these difference can be understood by noting: first, we have scaled the variance with the system size in eq. (4.3), namely $\sigma \to \frac{\sigma}{S}$; and second, for simplicity, we consider a symmetric $\mathbf{A}$, which follows Wigner's Semicircle law rather than Girko's Circular law. This difference contributes a prefactor of 4.

In contrast, in the LV dynamics we allow species to go extinct and some fraction $\phi_N = S^*/S$ of the species survive at the fixed point. Fig. 4.3 compares the spectrums of the full interaction matrix $\mathbf{A}$ and spectrum of the interaction matrix restricted to the surviving species $\mathbf{A}^*$. In the UFP phase, both of these spectrums follow Wigner Semicircle law but with slightly different radii: $2\sigma^2$ for $\mathbf{A}$, and $2\phi_N\sigma^2$ for $\mathbf{A}^*$, respectively. This small difference can result in big qualitative differences. The reason is that how many and which species survive in the LV dynamics can show big fluctuation in the MA and UG phases. For this reason, the interaction matrices may be not "self-averaging" in the large N limit, and consequently the spectrum of A^* does not converge to a deterministic spectrum.

The discussion above suggests an intuitive criteria for the stability of a fixed point of Lotka–Volterra dynamics where a fraction ϕ_N species survive, namely we should replace σ^2 by $\phi_N\sigma^2$ in May's stability criteria:

$$4\phi_N\sigma^2 = 1, \quad \phi_N = S^*/S. \tag{4.16}$$

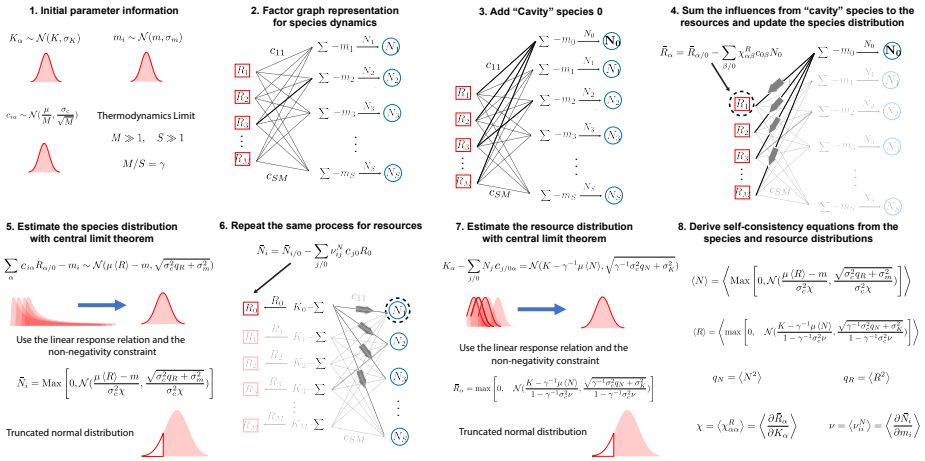

FIGURE 4.4: Schematic outlining steps in cavity solution. **1.** The initial parameter information consists of the probability distributions for the mechanistic parameters: K_α, m_i and $C_{i\alpha}$. We assume they can be described by their first and second moments. **2.** The species dynamics $N_i(\sum_\alpha c_{i\alpha}R_\alpha - m_i)$ in eqs. (4.18) are expressed as a factor graph. **3.** Add the "Cavity" species 0 as the perturbation. **4.** Sum the resource abundance perturbations from the "Cavity" species 0 at steady state and update the species abundance distribution to reflect the new steady state. **5.** Employing the central limit theorem and the non-negativity constraint, the species distribution is expressed as a truncated normal distribution. **6.** Repeat **Step 2-4** for the resources. **7.** The resource distribution is also expressed as a truncated normal distribution. **8.** The self-consistency equations are obtained from the species and resource distributions.

This equation was first derived in [Bun17, BBC18b] using the cavity method, replica method and random matrix theory. Alternatively, it can be understood by solving the self-consistency equation for the susceptibility ν:

$$\nu = \frac{1}{2\sigma^2}(1 - \sqrt{1 - 4\rho\phi_N\sigma^2}, \tag{4.17}$$

which has no real solution when $4\rho\phi_N\sigma^2 > 1$ for $\rho = 1$. This criteria also corresponds the point where the minimum eigenvalue λ_{min} first crosses zero. This suggests the susceptibility in the cavity equations have a deep connections with the spectrum of the random matrix describing interactions. This connections was developed in [CMIM19]. Finally, we note that for $\rho < 1$, we need to analyze the susceptibility in a complex plane. For brevity, we will not discuss these results here.

4.2 Cavity method for MacArthur's consumer-resource model

The MacArthur consumer resource dynamics is described with eq. (4.18) and eq. (4.19),

$$\frac{dN_i}{dt} = N_i(\sum_\beta c_{i\beta}R_\beta - m_i), \tag{4.18}$$

$$\frac{dR_\alpha}{dt} = R_\alpha(K_\alpha - R_\alpha - \sum_j N_j c_{j\alpha}) \tag{4.19}$$

This model can also be analyzed using the cavity method and in this section, we closely follow the derivation in [ABM18b]

As shown in Fig. 4.4: **step 1**, consumer preference $c_{i\alpha}$ are random variables drawn from a Gaussian distribution with mean μ/M and variance σ_c^2/M. It is helpful to decompose the consumer preference into an average and fluctuating component: $c_{i\alpha} = \mu/M + \sigma_c d_{i\alpha}$, where the fluctuating part $d_{i\alpha}$ obeys

$$\langle d_{i\alpha}\rangle = 0 \tag{4.20}$$

$$\langle d_{i\alpha}d_{j\beta}\rangle = \frac{\delta_{ij}\delta_{\alpha\beta}}{M}. \tag{4.21}$$

We also assume that both the carrying capacity K_α and the minimum maintenance cost m_i are independent Gaussian random variables with mean and covariance given by

$$\langle K_\alpha\rangle = K \tag{4.22}$$

$$\mathrm{Cov}(K_\alpha, K_\beta) = \delta_{\alpha\beta}\sigma_K^2 \tag{4.23}$$

$$\langle m_i\rangle = m \tag{4.24}$$

$$\mathrm{Cov}(m_i, m_j) = \delta_{ij}\sigma_m^2 \tag{4.25}$$

Let $\langle R\rangle = \frac{1}{M}\sum_\beta R_\beta$ and $\langle N\rangle = \frac{1}{S}\sum_j N_j$ be the average resource and average species abundance, respectively. With all these defined, we can re-write eq. (4.18) and eq. (4.19) as

$$\frac{dN_i}{dt} = N_i(\mu\langle R\rangle - m + \sum_\beta \sigma_c d_{i\beta}R_\beta - \delta m_i) \tag{4.26}$$

$$\frac{dR_\alpha}{dt} = R_\alpha(K + \delta K_\alpha - R_\alpha - \gamma^{-1}\mu\langle N\rangle - \sum_j \sigma_c d_{j\alpha}N_j) \tag{4.27}$$

where $\delta K_\alpha = K_\alpha - K, \delta m_i = m_i - m$ and $\gamma = M/S$. The basic idea of cavity method is to relate an ecosystem with $M+1$ resources (variables) and $S+1$ species to that with M resources and S species. In Fig. 4.4: **step 2**, we can express eq. (4.27) as a bipartite factor graph model for visualization. At **step 3**, we add a "cavity" species N_0 and a

"cavity" resource R_0 into the ecosystem,

$$\frac{dN_0}{dt} = N_0(\mu \langle R \rangle - m + \sum_\beta \sigma_c d_{0\beta} R_\beta - \delta m_0) \tag{4.28}$$

$$\frac{dR_0}{dt} = R_0(K + \delta K_0 - R_0 - \gamma^{-1}\mu \langle N \rangle - \sum_j \sigma_c d_{j0} N_j) \tag{4.29}$$

Adding new species and resource will perturb the original steady state. To characterize the perturbations, we introduce the following susceptibility matrices:

$$\chi^R_{\alpha\beta} = \frac{\partial \bar{R}_\alpha}{\partial K_\beta}, \qquad \chi^N_{i\alpha} = -\frac{\partial \bar{N}_i}{\partial K_\alpha}, \tag{4.30}$$

$$\nu^R_{\alpha i} = \frac{\partial \bar{R}_\alpha}{\partial m_i}, \qquad \nu^N_{ij} = \frac{\partial \bar{N}_i}{\partial m_j}. \tag{4.31}$$

We can express the steady-state species and resource abundances in the $(S+1, M+1)$ system with a first-order Taylor expansion around the (S, M) values. Because the mean part of the consumer resources, $\frac{\mu}{M} \sim \mathcal{O}(\frac{1}{M})$, are much smaller than the fluctuation term, $\sigma_c d_{i0} \sim \mathcal{O}(\frac{1}{\sqrt{M}})$, we can neglect the means and consider the perturbations due to the fluctuating components of the consumer preferences $\sigma_c d_{i0} R_0$ in eq. (4.27) and $\sigma_c d_{0\alpha} N_0$ in eq. (4.26) to m_i, and K_α, respectively.

Let us denote the species and resource abundances before adding the new species and resources by $N_{i/0}$ and $R_{\alpha/0}$ respectively, From the definition of the susceptibilities, we can relate the species and resources abundances after the perturbation (N_i and R_α) to the abundances before the perturbation ($N_{i/0}$ and $R_{\alpha/0}$) through the expressions:

$$\bar{N}_i = \bar{N}_{i/0} - \sigma_c \sum_{\beta/0} \chi^N_{i\beta} d_{0\beta} \bar{N}_0 - \sigma_c \sum_{j/0} \nu^N_{ij} d_{j0} \bar{R}_0 \tag{4.32}$$

$$\bar{R}_\alpha = \bar{R}_{\alpha/0} - \sigma_c \sum_{\beta/0} \chi^R_{\alpha\beta} d_{0\beta} \bar{N}_0 - \sigma_c \sum_{j/0} \nu^R_{\alpha j} d_{j0} \bar{R}_0 \tag{4.33}$$

Note $\sum_{j/0}$ and $\sum_{\beta/0}$ mean the sum excludes the new species 0 and the new resource 0. The next step is to plug eq. (4.32) and eq. (4.33) into eq. (4.28) and eq. (4.29) and solve for the steady-state value of N_0 and R_0.

4.2.1 Self-consistency equations for species

For the new "cavity" species N_0, the steady equation takes the form

$$\begin{aligned} 0 \quad &= \bar{N}_0(\mu \langle R \rangle - m - \delta m_0 - \sigma_c^2 \bar{N}_0 \sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} d_{0\alpha} d_{0\beta} \\ &\quad -\sigma_c^2 \bar{R}_0 \sum_{\beta/0, j/0} \nu^R_{\beta j} d_{0\beta} d_{0j} + \sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} + \sigma_c d_{00} \bar{R}_0) \end{aligned} \tag{4.34}$$

Notice that each of the sums in this equation is the sum over a large number of weak correlated random variables, and can therefore be well approximated by Gaussian random variables for large enough M and S. Using Eq. 4.20, we can calculate the sum of the random variables in the thermodynamic limit:

$$\sum_{\beta/0, j/0} \nu^R_{\beta j} d_{0\beta} d_{0j} = \frac{1}{M} \sum_{\beta/0, j/0} \nu^R_{\beta j} \delta_{j0} \delta_{\beta 0} = 0 \tag{4.35}$$

$$\sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} d_{0\alpha} d_{0\beta} = \frac{1}{M} \sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} \delta_{\alpha\beta} = \chi \tag{4.36}$$

where $\chi = \frac{1}{M} \sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} \delta_{\alpha\beta} = \frac{1}{M} \mathrm{Tr}(\chi^R_{\alpha\beta})$ is the average susceptibility. Using these observations about above sums, we obtain

$$\begin{aligned} 0 \quad &= \quad \bar{N}_0(\mu \langle R \rangle - m - \sigma_c^2 \chi \bar{N}_0 + \sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} \\ &+ \quad \sigma_c d_{00} \bar{R}_0 - \delta m_0) + \mathcal{O}(M^{-1/2}), \end{aligned} \tag{4.37}$$

Employing the Central Limit Theorem, we introduce an auxiliary Gaussian variable z_N with zero mean and unit variance and rewrite this as

$$\sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} + \sigma_c d_{0\beta} \bar{R}_0 - \delta m_0 = z_N \sqrt{\sigma_c^2 q_R + \sigma_m^2},$$

where q_R is the second moment of the resource distribution,

$$q_R = \langle R_\alpha^2 \rangle = \frac{1}{M} \sum_\beta R_\beta^2.$$

We can solve eq. (4.37) in terms of the quantities just defined:

$$\mu \langle R \rangle - m - \sigma_c^2 \chi \bar{N}_0 + \sqrt{\sigma_c^2 q_R + \sigma_m^2} z_N \leq 0 \tag{4.38}$$

Inverting this equation one gets the steady state of species

$$\bar{N}_0 = \max\left(0, \quad \frac{\mu\langle R\rangle - m + \sqrt{\sigma_c^2 q_R + \sigma_m^2} z_N}{\sigma_c^2 \chi}\right) \tag{4.39}$$

which is a truncated Gaussian.

Combining eq. (4.39) and eq. (4.11), we can easily write down the self-consistency equations for the fraction of non-zero species as well as the moments of their abundances at the steady state:

$$\phi_N = \frac{S^*}{S} = w_0\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right) \tag{4.40}$$

$$\langle N\rangle = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi}\right) w_1\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right) \tag{4.41}$$

$$q_N = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi}\right)^2 w_2\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right) \tag{4.42}$$

4.2.2 Self-consistency equations for resource

We now derive the equations for the steady-state of the resource dynamics. Inserting eq. (4.33) into eq. (4.29) gives:

$$\begin{aligned} 0 = \bar{R}_0(K - R_0 - \gamma^{-1}\mu\langle N\rangle + \sigma_c^2 \bar{N}_0 \sum_{\beta/0, j/0} \chi_{j\beta}^N d_{j0} d_{0\beta} \\ + \sigma_c^2 \bar{R}_0 \sum_{i/0, j/0} \nu_{ij}^N d_{0i} d_{0j} - \sum_{j/0} \sigma_c d_{j0} \bar{N}_{j/0} - \sigma_c d_{00} \bar{N}_0 + \delta K_0) \end{aligned} \tag{4.43}$$

where $\gamma = \frac{S}{M}$. We can also define the trace of the species susceptibility

$$\nu = \frac{1}{S} \sum_{i/0, j/0} \nu_{ij}^N \delta_{ij} = \frac{1}{S} \mathrm{Tr}(\nu_{ij}^N)$$

. Using the properties of $d_{i\alpha}$, i.e. eq. (4.20) and eq. (4.21), and following steps analogous to the derivation of eq. (4.36) and eq. (4.35), we get

$$\sum_{j/0} \sigma_c d_{j0} \bar{N}_{j/0} - \sigma_c d_{00} \bar{N}_0 + \delta K_0 = z_R \sqrt{\sigma_c^2 \gamma^{-1} q_N + \sigma_K^2},$$

where $q_N = \langle N_i^2\rangle = \frac{1}{S}\sum_j N_j^2$ and we have introduced an auxiliary Gaussian variable z_R with zero mean and unit variance. Plugging these expressions into Eq. 4.43 and solving for R_0 shows that the resource abundance distribution is also truncated gaussian

distribution of the form:

$$R_0 \quad = \quad \max\left(0, \quad \frac{K - \gamma^{-1}\mu \langle N\rangle + z_R}{1-\gamma^{-1}\sigma_c^2\nu}\right) \tag{4.44}$$

By analogy with the LV equations, We can easily write down the self-consistency equations for the fraction of non-zero resources as well as the moments of their abundances at the steady state:

$$\phi_R = \frac{M^*}{M} = w_0\left(\frac{\kappa - \gamma^{-1}\mu\langle N\rangle}{\sigma_{z_R}}\right) \tag{4.45}$$

$$\langle R\rangle = \left(\frac{\sqrt{\sigma_c^2\gamma^{-1}q_N + \sigma_K^2}}{1-\gamma^{-1}\sigma_c^2\nu}\right) w_1\left(\frac{K-\gamma^{-1}\mu\langle N\rangle}{\sqrt{\sigma_c^2\gamma^{-1}q_N+\sigma_K^2}}\right) \tag{4.46}$$

$$q_R = \left(\frac{\sqrt{\sigma_c^2\gamma^{-1}q_N + \sigma_K^2}}{1-\gamma^{-1}\sigma_c^2\nu}\right)^2 w_2\left(\frac{K-\gamma^{-1}\mu\langle N\rangle}{\sqrt{\sigma_c^2\gamma^{-1}q_N+\sigma_K^2}}\right) \tag{4.47}$$

The susceptibilities are given by averaging ν_{ii}^N and $\chi_{\alpha\alpha}^R$,

$$\chi \quad = \quad \left\langle \frac{\partial R_\alpha}{\partial K_\alpha}\right\rangle = \left\langle \frac{\partial R_0}{\partial K_0}\right\rangle = \frac{\phi_R}{1-\gamma^{-1}\sigma_c^2\nu}, \tag{4.48}$$

$$\nu \quad = \quad \left\langle \frac{\partial N_i}{\partial m_i}\right\rangle = \left\langle \frac{\partial N_0}{\partial m_0}\right\rangle = -\frac{\phi_N}{\sigma_c^2\chi} \tag{4.49}$$

Solving above two equations yields

$$\chi = \phi_R - \gamma^{-1}\phi_N, \quad \nu = -\frac{1}{\sigma_c^2}\frac{\phi_N}{\phi_R - \gamma^{-1}\phi_N}. \tag{4.50}$$

4.2.3 Comparison with numerics

Fig. 4.5 shows a comparison between the cavity solution and 1000 independent numerical simulations for various ecosystem properties such as the fraction of surviving species S^*/S, the fraction of surviving resources M^*/M, and the first and second moment of the species and resource distributions. As can be seen in the figure, our analytic expressions agree remarkably well over a large range of σ_c. As a further check on our analytic solution, we ran simulations where the $c_{i\alpha}$ were drawn from different distributions. One pathology of choosing $c_{i\alpha}$ from a Gaussian distribution is that when σ_c is large, many of consumption coefficients are negative. To test whether our cavity solution still describes ecosystems when $c_{i\alpha}$ are strictly positive, we compare our cavity solution to simulations where the $c_{i\alpha}$ are drawn from a binomial or uniform distribution in Fig. 4.6. As before, there is remarkable agreement between theoretical predictions and numerical simulations

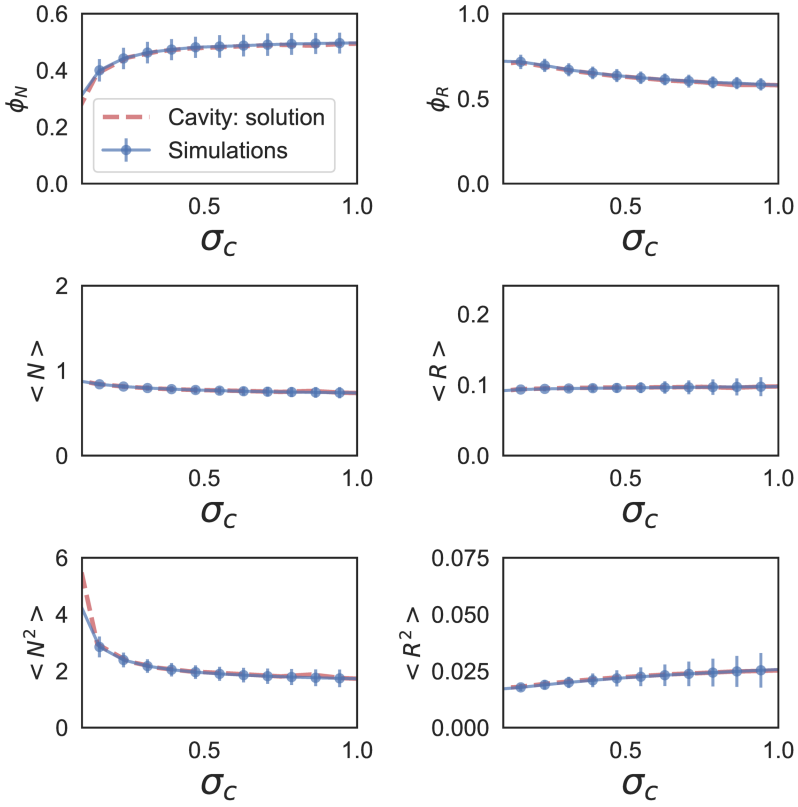

FIGURE 4.5: Comparison between cavity solutions (see main text for definition) and simulations for the fraction of surviving species $\phi_N = \frac{S^*}{S}$, the fraction of surviving species $\phi_R = \frac{M^*}{M}$ and the first and second moments of the species and resources distributions as a function of σ_c. The error bar shows the standard deviation from 100 numerical simulations with $M = S = 100$ $\mu = 1.$, $K = 1.$, $\sigma_K = 0.1$, $m = 0.1$, σ_m=0.01. Simulations were run using the CVXPY package [AVDB18].

in most of the range. Note that for the binomial distribution, if p is too small, the matrix become sparse and Gaussian distribution is not a good approximation and our cavity equations are less accurate.

4.2.4 Susceptibilities and Marchenko–Pastur distribution

A remarkable feature of the cavity solution is that we can directly relate the susceptibilities defined in eq. (4.30) and eq. (4.30) to results in Random Matrix (RMT) theory Recall, that these four susceptibility matrices that measure how the steady-state resource and species abundances respond to changes in the resource supply and species death(growth) rates. In fact, it turns out this relationship between susceptibilities and RMT is quite general and suggests a deep connection between RMT and phase transitions.

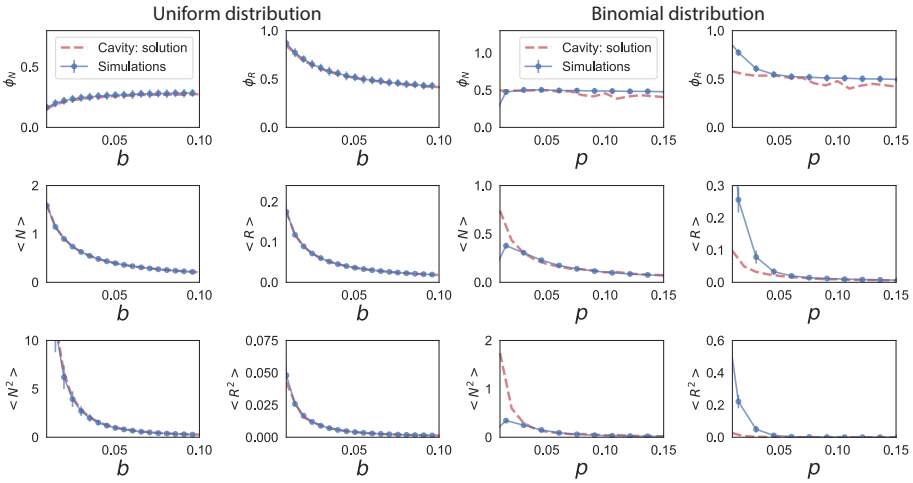

FIGURE 4.6: Comparison between cavity solutions and simulations for strictly positive distributions. The parameters are the same as Fig. 4.2 except $c_{i\alpha}$ is sampled from uniform distribution between 0 and b, and binomial distribution with nonzero probability p.

In what follows, we restrict the susceptibility matrices to surviving species and resources. The reason for this is that for the extinct species and resources, by definition the susceptibilities are zero. To proceed, we derive explicit equation satisfied by our four susceptibility matrices. Our starting point are the steady-state equations (Eq. 4.18) Eq. 4.19) for MacArthur's consumer-resource model:

$$0 = N_i(\sum_\alpha c_{i\alpha}R_\alpha - m_i), \tag{4.51}$$

$$0 = R_\alpha(K_\alpha - R_\alpha - \sum_j N_j c_{j\alpha}) \tag{4.52}$$

. Differentiating these equations with respect to K_β and m_j yields the relations

$$\begin{aligned} 0 &= \sum_{\alpha\in\mathbf{M}^*} c_{i\alpha}\chi^R_{\alpha\beta}, & \delta_{\alpha\beta} &= \chi^R_{\alpha\beta} + \sum_{j\in\mathbf{S}^*}\chi^N_{j\beta}c_{j\alpha} \\ \delta_{ij} &= \sum_{\alpha\in\mathbf{M}^*} c_{i\alpha}\nu^R_{\alpha j}, & 0 &= \nu^R_{\alpha i} + \sum_{j\in\mathbf{S}^*}\nu^N_{ji}c_{j\alpha}. \end{aligned} \tag{4.53}$$

where $\mathbf{M}^*$ and $\mathbf{S}^*$ denote the sets of surviving resources and species, respectively. These two equations can be written as single matrix equation for block matrices:

$$\begin{pmatrix} \mathbf{c} & 0 \\ \mathbb{1} & \mathbf{c}^T \end{pmatrix}\begin{pmatrix} \nu^R & \chi^R \\ \nu^N & \chi^N \end{pmatrix} = \mathbb{1} \tag{4.54}$$

To solve this equation, we define a $S^* \times S^*$ matrix: $A_{ij} = \sum_{\alpha \in M^*} c_{i\alpha} \bar{c}^T_{\alpha j}$. A straightforward calculation yields

$$\chi^R_{\alpha\beta} = \delta_{\alpha\beta} - \sum_{i\in \mathbf{S}^*} \sum_{j \in \mathbf{S}^*} c^T_{\alpha i} A^{-1}_{ij} c_{j\beta} \tag{4.55}$$

$$\chi^N_{i\alpha} = \sum_{j\in \mathbf{S}^*} A^{-1}_{ij} c_{j\beta}, \quad \nu^R_{\alpha i} = \sum_{j\in \mathbf{S}^*} c^T_{\alpha j} A^{-1}_{ji} \tag{4.56}$$

$$\nu^N_{ij} = -A^{-1}_{ij}, \quad i,j \in \mathbf{S}^* \text{ and } \alpha, \beta \in \mathbf{M}^* \tag{4.57}$$

Since the consumer preferences $c_{i\alpha}$ are random matrices, this suggests that we should be able to derive susceptibilities in cavity methods with Random Matrix Theory (RMT). We now show that this is indeed the case. Our starting point are the average susceptibilities which are defined as:

$$\chi = \frac{1}{M} \sum_{\alpha \in \mathbf{M}^*} \chi^R_{\alpha\alpha}, \quad \nu = \frac{1}{S} \sum_{i \in \mathbf{S}^*} \nu^N_{ii}.$$

From the cavity calculations, we only care about $\chi^R_{\alpha\beta}$ and ν^N_{ij}, because the other susceptibilities are lower order in $1/M$. We can combine with (4.56) to obtain

$$\begin{aligned} \chi &= \frac{1}{M}\mathrm{Tr}(\delta_{\alpha\beta}) - \frac{1}{M}\mathrm{Tr}\left(\sum_{i\in \mathbf{S}^*}\sum_{j\in \mathbf{S}^*} \bar{C}^T_{\alpha i} A^{-1}_{ij} \bar{C}_{j\beta}\right) \\ &= \frac{M^*}{M} - \frac{1}{M}\mathrm{Tr}\left(\sum_{i\in \mathbf{S}^*}\sum_{j\in \mathbf{S}^*} A^{-1}_{ij} \bar{C}_{j\beta} \bar{C}^T_{\beta h}\right) \\ &= \frac{M^*}{M} - \frac{S^*}{M} = \phi_R - \gamma^{-1}\phi_N \end{aligned} \tag{4.58}$$

A_{ij} is the outer product of a random matrix $\mathbf{c}$ with itself, i.e., a Wishart matrix. The underlying reason for this is the bipartite nature of the consumer resource models resulting from the presence of two types of degrees of freedom: resources and species [RCKT08, RCKT08, AF19, AABF19]. Random Wishart matrices are well-known to follow a different eigenvalue distribution, the Marchenko-Pastur law[MP67b] given by

$$\rho(x) = \frac{1}{2\pi\sigma^2_c c x}\sqrt{(b-x)(x-a)} + \Theta(c-1)(1-c^{-1})\delta(x) \tag{4.59}$$

where $c = \frac{S^*}{M^*}$ and $\Theta(x)$ represents the Heaviside step function. Since $S^* < M^*$ is always true, the second term can be ignored. $A_{ij} = \sum_{\alpha \in \mathbf{S}^*} \bar{C}_{i\alpha} \bar{C}^T_{\alpha j}$ takes the form of a Wishart Matrix. We will exploit this to calculate χ and ν. Notice,

$$\nu = = -\frac{1}{S}\mathrm{Tr}(A^{-1}_{ij}) = -\frac{1}{S}\sum_{i=1}^{S^*} \lambda_i^{-1} \tag{4.60}$$

where λ_i is the eigenvalue of A_{ij}.

Substituting equation (4.59) into the expression for ν and replacing the sum with an integral yields:

$$\begin{aligned} \nu &= -\frac{S^*}{S}\int_a^b \frac{1}{x}\rho(x)dx = -\frac{S^*}{S}\frac{a+b-2\sqrt{ab}}{4\sigma_c^2 y\sqrt{ab}} \\ &= -\frac{1}{\sigma_c^2}\frac{\phi_N}{\phi_R - \gamma^{-1}\phi_N} \end{aligned} \tag{4.61}$$

The second line of equation (4.61) is obtained by transferring the integral function to a complex analytic function and applying the residue theorem.

Chapter 5

When will complex ecosystems behave like random systems?

In 1972, Robert May triggered a worldwide research program studying ecological communities using random matrix theory. Yet, it remains unclear if and when we can treat real communities as random ecosystems. In Chapter 2, we have shown that such models, initialized with random parameters, can predict lab experiments on complex microbial communities [GI83, MCM20] and reproduce large-scale ecological patterns observed in field surveys, including the Earth and Human Microbiome Projects [MCM20]. This suggests that the large-scale, reproducible patterns we see across Microbiomes are emergent features of random ecosystems.

Yet, it remains unclear why random ecosystems can accurately describe real ecological communities. To answer these questions, in this paper we exploit ideas from random matrix theory and statistical physics to analyze generalized consumer-resource models in spirit of May's original analysis. We show that the macroscopic ecological properties of diverse ecosystems can be described using random ecosystems, much like thermodynamic quantities like pressure and average energy of the ideal gas can be described by considering particles to be random and independent.

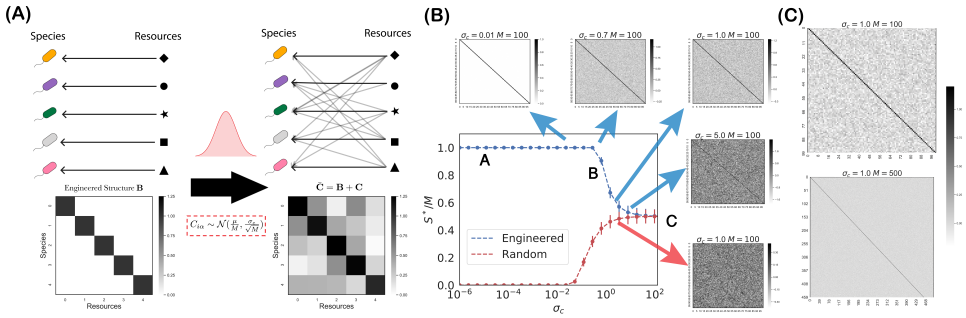

FIGURE 5.1: **Random interactions destabilize an ecosystem of specialist consumers.** **(A)** Left: an ecosystem with system size $M = 5$ starts with specialists consuming only one type of resource, resulting in a consumer preference matrix $\mathbf{B} = \mathbb{1}$. Right: off-target consumption coefficients $\mathbf{C} \sim \mathcal{N}(\frac{\mu}{M}, \frac{\sigma_c}{\sqrt{M}})$ are sampled from a Gaussian distribution, resulting in an overall consumer preference matrix $\bar{\mathbf{C}} = \mathbf{B} + \mathbf{C}$. **(B)** Fraction of surviving species S^*/M vs. σ_c, numerically computed using $M = 100$ for an ecosystem described by Eq. 5.2, along with the corresponding results for a completely random ecosystem with $\mathbf{B} = 0$. The error bar shows ± 1 standard deviation from 10000 independent realizations. Also shown are examples of the matrices $\bar{\mathbf{C}}$ employed in the simulations. **(C)** Heatmap for the identity matrix plus a gaussian random matrix with $\sigma_c = 1$ for two system sizes: $M = 100$ and $M = 500$.

5.1 Models

To explore these ideas, we devised a more concrete version of May's original thought experiment describing an ecosystem consisting of S non-interacting species where interactions are gradually turned on. May's original argument only considered the local dynamics near a pre-specified equilibrium point that eventually becomes unstable. Since we are interested in exploring what happens in consumer resource models, we must make additional modeling assumptions to arrive at a complete set of nonlinear dynamics. We focus on numerous variants of the Consumer Resource Model (CRM)[ML67a], including different choices of resource dynamics, consumer preferences, as well as more dramatic variants such as the Microbial Consumer Resource Model introduced in [GLB+18, MCGM20, MCGM20].

The original MacArthur Consumer Resource Model [ML67a] consists of S species or consumers with abundances N_i $(i = 1...S)$ that can consume one of M substitutable resources with abundances R_α $(\alpha = 1...M)$, whose dynamics are described by the equations

$$\begin{cases} \frac{dN_i}{dt} = N_i(\sum_\beta \bar{C}_{i\beta} R_\beta - m_i) \\ \frac{dR_\alpha}{dt} = R_\alpha(K_\alpha - R_\alpha - \sum_j N_j \bar{C}_{j\alpha}). \end{cases} \tag{5.1}$$

The consumption rate of species i for resource α is encoded by the entry $\bar{C}_{i\alpha}$ in the $S \times M$ consumer preference matrix $\bar{\mathbf{C}}$, K_α is the carrying capacity of resource α, and

m_i is a maintenance energy that encodes the minimum amount of energy that a species i must harvest from the environment in order to survive. When the system is in the steady state, some species and resources can vanish. We denote the numbers of surviving species and resources by S^* and M^*, respectively, and in general at steady state we will have $S^* \leq S$ and $M^* \leq M$. For this reason, we refer to this model as the CRM *with* resource depletion and consider its effects analytically and numerically in later sections.

In the beginning, we focus primarily on a popular variant of the original CRM introduced by Tilman with slightly different resource dynamics[Til82b]:

$$\begin{cases} \frac{dN_i}{dt} = N_i(\sum_\beta \bar{C}_{i\beta} R_\beta - m_i) \\ \frac{dR_\alpha}{dt} = K_\alpha - R_\alpha - \sum_j N_j \bar{C}_{j\alpha}. \end{cases} \tag{5.2}$$

The main difference between this model variant and the original CRM is that consumers can no longer deplete a resource (i.e. $\mathbf{M}^* = \mathbf{M}$). This makes this models significantly easier to analyze (especially within the context of Random Matrix Theory) and leads to much simpler analytic expressions. For this reason, we largely focus on this model *without* resource depletion. However, we note that a major drawback of this model is that it can lead to unphysical, negative resource concentrations and hence is physically flawed. Despite this limitation, the CRM without resource depletion captures almost all the qualitative behaviors present in more complicated and physically realistic CRMs (though there are some subtle but important differences discussed below).

Both the models in Eq. 5.1 and Eq. 5.2 make very specific assumptions about resource dynamics (i.e. that resources are themselves self-replicating entities that can be described by logistic growth in the absence of consumers). To check the generality of our results, we also numerically analyzed generalizations of the CRM including linear resource dynamics where resources are supplied externally, and a model of microbial ecology with trophic feedbacks where organisms can feed each other via metabolic byproducts [GLB+18, MICG+19, MCM20, MCGM20]. Furthermore, for simplicity, in most of this work we assume that $S = M$. However, we have numerically checked that our results are robust to breaking on this assumption.

In CRMs, the identity of each species is specified by its consumption preferences. In real ecosystems, it is well established that organisms can exhibit strong consumer preferences for particular resources. However, recent work has shown that consumer resource models with random consumer preferences can reproduce experimental observations in field surveys and laboratory experiments [GLB+18, MCM20]. To understand this phenomena, we asked how adding noise to consumer preferences changes macroscopic ecosystem level properties like diversity and average productivity. To do so, we considered a thought

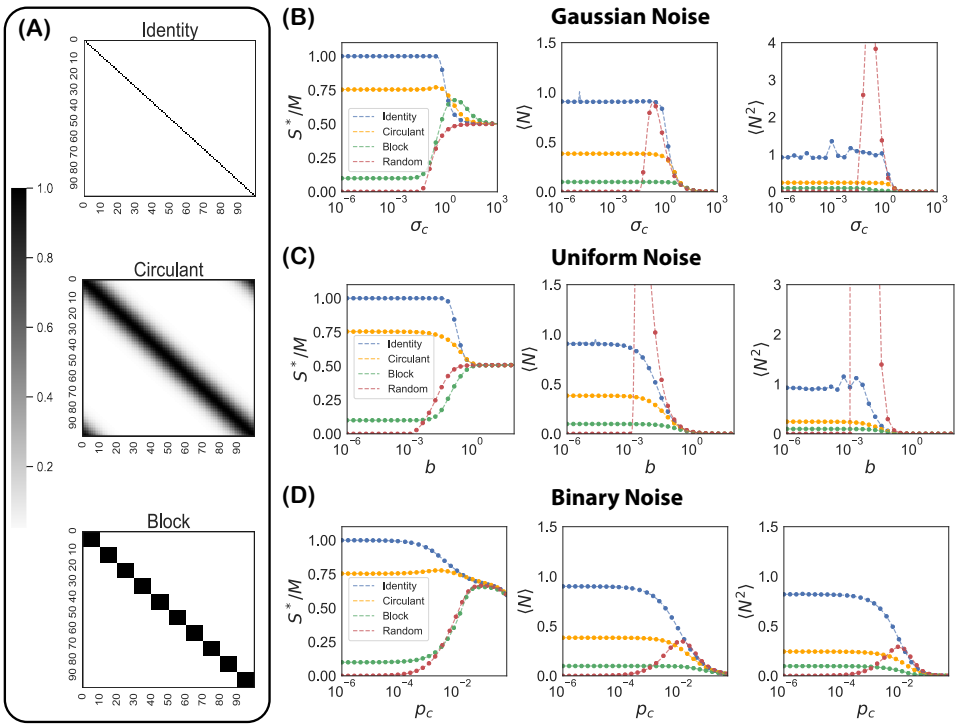

FIGURE 5.2: **Community properties for structured and random ecosystems. (A)**: Examples of designed interactions Top: the identity matrix; Middle: a Gaussian-type circulant matrix; Bottom: a block matrix (see Methods for details). Simulations of designed and random ecosystems where the random component of the the consumer preferences **C** are sampled from a **(B)** Gaussian distribution $\mathcal{N}(0, \frac{\sigma_c}{\sqrt{M}})$, **(C)** Uniform Distribution: $\mathcal{U}(0, b)$ or a **(D)**: Binomial distribution: $Bernoulli(p_c)$. The plots show the fraction of surviving species S^*/M, mean species abundance $\langle N \rangle$, and second moment of the species abundances $\langle N^2 \rangle$ for designed and purely random ecosystems the number of non-specific consumer preferences is increased.

experiment where we started with non-interacting species where each species consumes its own resource, and then added "noise" to the consumer resource preferences.

A set of non-interacting species can be constructed by engineering each species to consume a different resource type, with no overlap between consumption preferences. For example, one can imagine designing strains of *E. coli* where each strain expresses transporters only for a single carbon source with all other transporters edited out of the genome: i.e a strain that can only transport lactose, another strain that can only transport sucrose, etc. An ecosystem with such consumer preference structure is shown in Figure 5.1(A). In such an experiment, horizontal gene transfer would eventually begin distributing transporter genes from one strain to another, so a realistic model would have to allow for some amount of unintended, "off-target" resource consumption. In line with May, we can model the consumer preferences $\bar{C}_{i\alpha}$ of species i for resource α in such an ecosystem as the sum of the identity matrix $\mathbb{1}$ and a random component $C_{i\alpha}$

with variance σ^2 that encodes non-specific preferences (see Figure 5.1A right). In other words, the full consumer matrix can be written as $\bar{\mathbf{C}} = \mathbf{I} + \mathbf{C}$.

5.2 Phase transition to random ecosystems

Figure 5.1(B) shows how the number of surviving species at steady-state as one adds more and more non-specific resource preferences to an ecosystem initially composed of non-interacting species. Just as in May's analysis, the appropriate measure of the importance of the random component is the root-mean-squared off-target consumption $\sigma_c = \sqrt{M\sigma^2}$ (recall $M = S$). This scaling reflects the fact that two consumer matrices $\bar{\mathbf{C}}$ with the same σ_c but different system sizes M can have very different amounts of absolute noise as shown Figure 5.1(C), but exhibit almost identical community-level properties (with all differences coming from finite size effects). Figure 5.1(B) shows the fraction of surviving species S^*/M in the ecosystem as a function of σ_c. At small values of σ_c, all the species survive and $S^* = S$. As high as $\sigma_c = 0.7$, almost all of the original species are still present in the community. But between $\sigma_c = 0.7$ and $\sigma_c = 1$, there is a sharp transition in community structure, which results in about half of the original species becoming extinct.

Remarkably, the fraction of surviving species converges to the same value as for a completely random consumer preference matrix and remains finite as $\sigma_c \to \infty$ [SCG+18]. This means that ecosystems with an arbitrarily large number of species can be stably formed by considering a sufficiently large initial ecosystem. We also examined two other community-level properties: the mean species abundance $\langle N \rangle$ (i.e. the average productivity), and the second moment of the population size $\langle N^2 \rangle$, which includes information about the distribution of population sizes of various species. Figure 5.2 shows that both of these quantities are also well-approximated by the random consumer preference matrix for $\sigma_c > 1$. These numerical predictions are in excellent agreement with analytic predictions derived in the $S \to \infty$ limit derived in Section 5.3 using the cavity method [Bun17, ABM18b].

This convergence to random ecosystem behavior is quite robust, and holds for other choices of designed consumer preferences beyond the identity matrix considered above. Figure 5.2 shows numerical simulations of the diversity S^*/M, average productivity $\langle N \rangle$, and second moment of the species abundances $\langle N^2 \rangle$ as a function of the noise σ_c for two other choices of designed consumer preference matrices: a block structure with pre-defined groups of species exhibiting strong intra-group competition and a unimodal structure where each species is more likely to consume resources similar to its preferred resource. Once again, we see that the ecosystem quickly transitions to a behavior where

these macroscopic properties are indistinguishable from those of a random ecosystem. The primary effect of the choice of consumer preference matrix is to adjust the threshold value of σ_c where the transition to typicality takes place. In all cases, we find that the random behavior takes over when the average total off-target consumption capacity over all M resource types becomes greater than the consumption of the primary resource in the original designed ecosystem in the absence of noise, the same as May's stability criteria [May72].

The character of the self-organized state is also robust to changes in the sampling scheme for the random component of the consumer preferences. Gaussian noise in consumer preferences allows the clearest comparison to May's result but also sometimes results in non-physical negative values for consumer preferences. We therefore tested two sampling schemes that always produce positive values for consumer preferences: uniformly sampling the random component of preferences $C_{i\alpha}$ in an interval from 0 to b, and binary sampling where $C_{i\alpha} = 1$ with probability p_c and zero otherwise. Changing b or p_c affects both the mean and the variance of the random components of the consumer preferences simultaneously making it difficult to directly compare to the Gaussian case. Nonetheless, as can be seen in the Figure 5.2, the qualitative behaviors is identical to the Gaussian case, with macroscopic ecological properties becoming indistinguishable from those of a fully random ecosystem when the average off-target resource consumption comparable to the the consumption of the designed resources.

5.2.1 Sensitivity to perturbations and the transition to typicality

To better understand why mass extinctions happen at $\sigma_c^* \sim 1$ and allow for comparison with May's original analysis, we calculated an effective species-species competition matrix A_{ij} between species for an ecosystem whose dynamics are governed by Eq. 5.2. We exploited the observation by MacArthur and others that if resource abundances always remain close to their steady state values, the steady-states of the CRM coincide with those of an effective generalized Lotka-Volterra model of the form

$$\frac{dN_i}{dt} = N_i \left(\sum_{\alpha \in \mathbf{M}} C_{i\alpha} K_\alpha - m_i - \sum_j A_{ij} N_j \right), \tag{5.3}$$

with the species-species interaction matrix given by

$$A_{ij} = \sum_{\alpha \in \mathbf{M}} \bar{C}_{i\alpha} \bar{C}^T_{\alpha j} \tag{5.4}$$

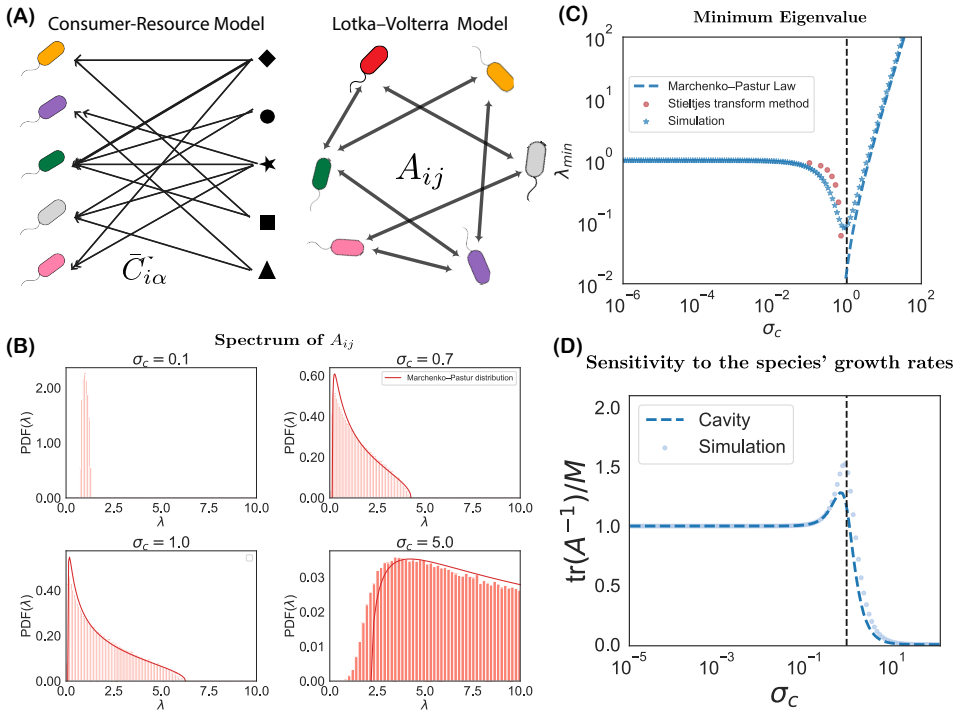

FIGURE 5.3: **Effect of random interactions on ecosystem sensitivity**. **(A)**: The bipartite interactions $\bar{C}_{i\alpha}$ in MacArthur's consumer-resource model can be mapped to pairwise competition coefficients A_{ij} in generalized Lotka-Volterra equations through $A_{ij} = \sum_{\alpha \in \mathbf{M}} \bar{C}_{i\alpha} \bar{C}^T_{\alpha j}$. **(B)** Spectra of A_{ij} at different σ_c for $\bar{\mathbf{C}} = \mathbb{1} + \mathbf{C}$, where $\mathbf{C}$ is a random matrix with i.i.d entries drawn from a normal distribution with mean zero and standard deviation σ_c. The red solid line is the Marchenko-Pastur distribution. **(C)**: Comparison between numerical simulations and analytic results for the minimum eigenvalue of $\mathbf{A}$ at different σ_c. **(D)**: Comparison between numerical simulations and analytic solutions for the mean sensitivity ν of steady-state population sizes to changes in species growth rates.

(see Figure 5.3(A) for details). This matrix is related to May's community matrix governing stability $\mathbf{J}$ discussed in the introduction through the relation $J_{ij} = -\bar{N}_i A_{ij}$, where $\bar{N}_i$ is the steady-state abundance of species i. For symmetric interaction matrices of the form in Eq. 5.4, it is possible to prove that the largest eigenvalue $\lambda_{\max}$ of $\mathbf{J}$ reaches zero from below only when the smallest eigenvalue $\lambda_{\min}$ of $\mathbf{A}$ reaches zero from above (see Appendix A).

As shown in Figure 5.1(B), the behavior broadly falls into one of three different regimes depending on the amount of noise introduced in the consumer preferences: a low-noise regime when $\sigma_c \ll 1$, a cross-over regime when $0 \ll \sigma_c \leq 1$, and a high-noise regime when $\sigma_c > 1$. Figure 5.3(B) shows how the eigenvalue spectrum of the corresponding Lotka-Volterra interaction matrix $\mathbf{A}$ change as σ_c increases.

Low-noise regime ($\sigma_c \ll 1$): In the low-noise regime, the engineered structure in

the consumer preference controls large scale ecological properties. Furthermore, the eigenvalue spectrum of the LV-interaction matrix $\mathbf{A}$ is centered around 1 reflecting the fact there is very little competition between species (i.e. species still occupy largely independent niches). For this reason, in this regime all the initial species in the ecosystem survive to steady-state so that $S^*/M = 1$.

Crossover regime ($0 \ll \sigma_c \leq 1$): With increasing σ_c, the eigenvalues due the noise component in $\mathbf{A}$ repel each other like in the Coulomb gas and the spectrum spreads out [Dys62]. $\lambda_{\min}$ decreases until it reaches the threshold of stability $\lambda_{\min} \cong 0$ at $\sigma_c^* \approx 1$. Note that $\lambda_{\min}$ is close to 0 but not exactly at 0 because the steady-state of the CRM is alway stable [Che90b]. In this regime even a small environmental perturbations or small amounts of demographic noise can result in species extinctions [DB20]. This is closely related to the divergence of structural stability when $\lambda_{\min} \sim 0$[RSB14]. In Section 5.3 we show analytically using the Cavity method [Bun17, ABM18b] that in the limit $M \to \infty$, $\lambda_{\min}$ is approaches 0 from above when $\sigma_c^* = 1$. At $\sigma_c \sim 1$ the engineered structure and noise have comparable amplitudes. For the case where the consumer preferences are chose to be binary noise, this threshold corresponds to a critical noise level $p_c \sim \frac{1}{M}$, meaning on average there is one random nonzero element in the row besides the diagonal one. More generally, our numerics suggest that the threshold to typicality occurs in a wide variety of models when the expected off-target resource consumption rates become comparable to the the consumption rate for the designed resources.

Noise-dominated regime ($\sigma_c > 1$) In this regime, we observe two new phenomena that were not accessible in May's original framework. First, the spectrum of the species-species interaction matrix A_{ij} approaches the Marchenko-Pastur law [MP67b],

$$\rho(x) = \frac{1}{2\pi\sigma_c^2 cx}\sqrt{(b-x)(x-a)} + \Theta(c-1)(1-c^{-1})\delta(x) \tag{5.5}$$

where $a = \sigma_c^2(1-\sqrt{c})^2$, $b = \sigma_c^2(1+\sqrt{c})^2$, $c = S^*/M$ and $\Theta(x)$ represents the Heaviside step function. This differs from May's analysis where the spectrum of the interaction network follows Girko's Circular law [RCKT08, AF19, AABF19]. The reason for this difference is that species-species interaction matrix obtained from the CRM is the outer product of a random matrix $\bar{\mathbf{C}}$ with itself (i.e., a Wishart matrix, see Eq. 5.4), reflecting the fact that the CRM has two different kinds of degrees of freedom: resources and species. The Marchenko-Pastur law is the distribution we would expect for an ecosystem with completely random consumer preferences [MP67b]. This helps explain our earlier observations that community-level observables of ecosystems are indistinguishable from the purely random ecosystems when σ_c is sufficiently large (see Figure 5.3(B)).

Secondly, as σ_c increases past 1 and ecosystem properties become typical, the resulting ecosystems once again become insensitive to external perturbation [DB20]. To see this,

we note that we can measure sensitivity to perturbations by examining the minimum eigenvalue of the interaction matrix A_{ij}, with larger $\lambda_{\min}$ meaning decreased sensitivity to perturbations (see Appendix A.1). The minimum eigenvalue in the Marchenko-Pastur Distribution is located at

$$\lambda_{\min} = \sigma_c^2(1 - \sqrt{S^*/M})^2. \tag{5.6}$$

As one increases σ_c, $S^*/M \to 1/2$ from above since there is increases competition between species for shared resources. Consequently, $\lambda_{\min}$ is always be much larger than zero once ecosystems crossover to their typical behavior.

The above analysis suggests that $\lambda_{\min}$ is an important property that can be used to characterize the three regimes seen in Figure 5.3(C). In the low-noise regime, species-species interactions are weak and $\lambda_{\min} \approx 1$, whereas in the high-noise regime $\lambda_{\min} = \sigma_c^2(1 - \sqrt{S^*/M})^2$. The calculation of $\lambda_{\min}$ in Regime B is challenging because of the mixture between the engineered structure and noise. However, we can use techniques from RMT for wireless communication (i.e information-plus-noise models) to analytically estimate $\lambda_{\min}$ [CD11, LV+11]. The results are shown in the red scatter points in Figure 5.3(D) (see Appendix 5.4.3). As discussed above, $\lambda_{\min}$ approaches zero as σ_c approaches one.

The spectrum of $\mathbf{A}$ also contains quantitative information about the sensitivity of the ecosystem in the Cavity method. Specifically, as shown in Section 5.3, we can define a susceptibility ν that measures the average response of the steady-state population size $\bar{N}_i$ to perturbing of the species maintenance cost m_i (see Eq. 5.2). We further show that ν is directly related to the the sum of the inverse eigenvalues of A_{ij} through the expression

$$\nu = \frac{1}{M}\sum_i (1/\lambda_i) = \frac{1}{M}\mathrm{tr}(\mathbf{A}^{-1}). \tag{5.7}$$

Figure 5.3(D) shows that this quantity is initially constant as σ_c is increased from 0, then reaches the maximum value at $\sigma_c = 1$, and finally rapidly decreases to near zero. In Section 5.3 we provide analytical calculations based on the cavity method confirming these numerical results.

Note that our results are not restricted to Gaussian noise but also apply to the other cases where the noise in consumer preferences is binary or uniform. This is because the *central limit theorem* guarantees that the statistics of eigenvalues of large random matrices converges to the statistics in Gaussian random matrices for many biologically plausible choices of consumer preferences.

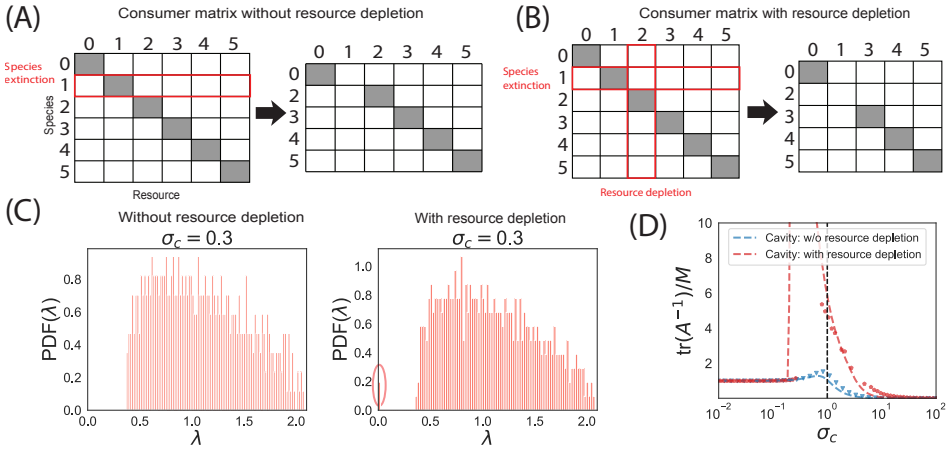

FIGURE 5.4: **Effect of resource extinction on an ecosystem**. A schematic for the consumer preference matrix with ((**A**)) and ((**B**))without resource extinction for specialist consumers that each eat independent resources. The left schematic corresponds to the initial consumer matrix, and the right schematic to the consumer matrix after species and resource extinctions. Notice that resource extinctions can result in singular consumer matrices **(C)** Spectra of A_{ij} at $\sigma_c = 0.3$ with consumer matrices chosen as in Figure 5.3 with (left) and without resource extinction (right). The zero modes are marked with a red ellipse. **(D)** the mean sensitivity ν of steady-state at different σ_c. The dashed lines in **(D)** are cavity solutions. The scatter points are results from numerical simulations. See Section 5.3 for detailed calculations.

5.2.2 Effect of resource depletion

Thus far we have focused on a CRM without resource extinctions specified by Eqs. 5.2. As discussed extensively in Section 5.3, if we instead allow for resource extinction (Eqs. 5.1), somewhat surprisingly, our cavity method predicts a first-order phase transition to typicality rather than a cross-over as is the case without resource extinction. The signature of such a first order transition is the divergence of the susceptibility matrix ν discussed above. Figure 5.4 shows ν with and without resource extinction, numerically confirming the existence of this first order transition. This first order transition is also reflected in the spectrum of the interaction matrix $\mathbf{A}$ through the the appearance of zero eigenvalue modes for CRMs when resources can go extinct.

The existence of zero modes can be understood by noting that resource extinction and species extinction correspond to the column and row deletion in the consumption matrix (shown in Figure 5.4(A)). Such deletions can change the engineered component of the effective consumer preferences for surviving species and resources, resulting in large fluctuations in the interaction matrix $\mathbf{A}$. In the presence of these large fluctuations, the interaction matrix no longer self-averages, giving rise to the observed first-order phase transition. This same mechanism also leads to a first-order phase transition to typical

behavior when the engineered portion of the consumer resources is block diagonal, even in the absence of resource extinctions (see Figure 5.4).

5.3 Cavity solution

When the designed component of the consumer preferences is the identity (i.e $\mathbf{B} = \mathbb{1}$, the effect of random off-target consumption on system-scale properties can be computed analytically in the $M, S \to \infty$ limit using the cavity method introduced in Chapter 4. The cavity calculation is straightforward but tedious. For this reason, it is helpful to introduce the notation as before:

- $\frac{M^*}{M} = \phi_R$, $\langle R \rangle = \frac{1}{M}\sum_\beta R_\beta$ and $q_R = \frac{1}{M}\sum_\beta R_\beta^2 = \langle R^2 \rangle$, where M^* is the number of surviving resources.
- $\frac{S^*}{S} = \phi_N$, $\langle N \rangle = \frac{1}{S}\sum_j N_j$ and $q_N = \frac{1}{S}\sum_j N_j^2 = \langle N^2 \rangle$, where S^* is the number of surviving species.
- $C_{i\alpha} \equiv \frac{\mu}{M} + \sigma_c d_{i\alpha}$ assuming $\langle d_{i\alpha} \rangle = 0$, $\langle d_{i\alpha} d_{j\beta} \rangle = \frac{\delta_{ij}\delta_{\alpha\beta}}{M}$. with $\langle c_{i\alpha} \rangle = \frac{\mu}{M}$, $\langle c_{i\alpha} c_{j\beta} \rangle = \frac{\sigma_c^2}{M}\delta_{ij}\delta_{\alpha\beta} + \frac{\mu^2}{M^2} \approx \frac{\sigma_c^2}{M}\delta_{ij}\delta_{\alpha\beta}$.
- $K_\alpha = K + \delta K_\alpha$ with $\langle K_\alpha \rangle = \frac{1}{M}\sum_\beta K_\beta = K$, $\langle \delta K_\alpha \delta K_\beta \rangle = \delta_{\alpha\beta}\sigma_K^2$.
- $m_i = m + \delta m_i$ with $\langle m_i \rangle = m$, $\langle \delta m_i \delta m_j \rangle = \delta_{ij}\sigma_m^2$.
- $\gamma = \frac{M}{S}$ and for the identity matrix $\gamma = 1$.

Following similar steps as in Chapter 4 and [ABM18b], we perturb the ecosystem with a new species and resource N_0 and R_0. Ignoring $\mathcal{O}(1/M)$ terms yields the following equations:

$$\frac{dN_i}{dt} = N_i\left[R_i - m + \sum_\beta (\frac{\mu}{M} + \sigma_c d_{i\beta})R_\beta + (\frac{\mu}{M} + \sigma_c d_{i0})R_0 - \delta m_i\right] \tag{5.8}$$

$$\frac{dR_\alpha}{dt} = R_\alpha\left[K + \delta K_\alpha - R_\alpha - N_\alpha - \sum_j (\frac{\mu}{M} + \sigma_c d_{j\alpha})N_j - (\frac{\mu}{M} + \sigma_c d_{0\alpha})N_0\right] \tag{5.9}$$

$$\frac{dN_0}{dt} = N_0\left[R_0 - m + \sum_\beta (\frac{\mu}{M} + \sigma_c d_{j\alpha})R_\beta - \delta m_0\right] \tag{5.10}$$

$$\frac{dR_0}{dt} = R_0\left[K + \delta K_0 - R_0 - N_0 - \sum_j (\frac{\mu}{S} + \sigma_c d_{j0})N_j\right] \tag{5.11}$$

Denote by $\bar{N}_{\alpha/0}$, $\bar{R}_{\alpha/0}$ and $\bar{N}_i$, $\bar{R}_\alpha$ the equilibrium values of the species and resources before and after adding the newcomers, respectively. These can be related to each other using the susceptibilities defined above:

$$\bar{N}_i = \bar{N}_{i/0} - \sigma_c \sum_j \nu_{ij}^N d_{j0} R_0 - \sigma_c \sum_\beta \chi_{i\beta}^N d_{0\beta} N_0 \tag{5.12}$$

$$\bar{R}_\alpha = \bar{R}_{\alpha/0} - \sigma_c \sum_i \nu_{\alpha i}^R d_{i0} R_0 - \sigma_c \sum_\beta \chi_{\alpha\beta}^R d_{0\beta} N_0 \tag{5.13}$$

In what follows we assume Replica Symmetry. In this case, the sums in the equations above can be approximated as Gaussian random variables. For this reason, it is helpful to introduce new auxiliary random variables:

$$z_N = \sum_\beta \sigma_c \bar{R}_{\beta/0} d_{0\beta} - \delta m_0 \tag{5.14}$$

$$z_R = \sum_j \sigma_c \bar{N}_{j/0} d_{j0} - \delta K_0 \tag{5.15}$$

where $\langle z_N \rangle = 0$, $\sigma_{z_N} = \sqrt{\sigma_c^2 q_R + \sigma_m^2}$ and $\langle z_R \rangle = 0$, $\sigma_{z_R} = \sqrt{\sigma_c^2 q_N + \sigma_K^2}$.

Case 1: both R_0 and N_0 are positive. Following calculations analogous to [ABM18b] and noting that $\gamma = \frac{M}{S} = 1$ yields:

$$\bar{R}_0 = \max\left[0, \frac{\sigma_c^2 \chi (K - \mu \langle N \rangle + z_R) - \mu \langle R \rangle + m - z_N}{(1 - \sigma_c^2 \nu)\sigma_c^2 \chi + 1}\right] \tag{5.16}$$

$$\bar{N}_0 = \max\left[0, \frac{(1 - \sigma_c^2 \nu)(\mu \langle R \rangle - m + z_N) + K - \mu \langle N \rangle + z_R}{(1 - \sigma_c^2 \nu)\sigma_c^2 \chi + 1}\right] \tag{5.17}$$

Case 2: either R_0 or N_0 is zero. We get exactly the same expression as the random ecosystem we derived in [ABM18b].

$$\bar{R}_0 = 0, \quad \bar{N}_0 = \frac{\mu \langle R \rangle - m + z_N}{\sigma_c^2 \chi} \quad \text{or,} \quad \bar{N}_0 = 0, \quad \bar{R}_0 = \frac{K - \mu \langle N \rangle + z_R}{1 - \sigma_c^2 \nu} \tag{5.18}$$

Case 3: both R_0 and N_0 are zero, namely,

$$\bar{R}_0 = 0 \text{ and } \bar{N}_0 = 0. \tag{5.19}$$

Combining the cases above, the steady state solution is a Gaussian mixture depending on the positivity of R_0 and N_0.

$$\begin{aligned}\bar{R}_0 = \Theta(R_0)\Theta(N_0)\frac{\sigma_c^2\chi(K-\mu\langle N\rangle + z_R) - \mu\langle R\rangle + m - z_N}{(1-\sigma_c^2\nu)\sigma_c^2\chi + 1} \\ + \Theta(R_0)(1-\Theta(N_0))\frac{K-\mu\langle N\rangle + z_R}{1-\sigma_c^2\nu}\end{aligned} \tag{5.20}$$

$$\begin{aligned}\bar{N}_0 = \Theta(N_0)\Theta(R_0)\frac{(1-\sigma_c^2\nu)(\mu\langle R\rangle - m + z_N) + K - \mu\langle N\rangle + z_R}{(1-\sigma_c^2\nu)\sigma_c^2\chi + 1} \\ + \Theta(N_0)(1-\Theta(R_0))\frac{\mu\langle R\rangle - m + z_N}{\sigma_c^2\chi}\end{aligned} \tag{5.21}$$

Cavity equations for the susceptibilities can be obtained directly by differentiating these equations:

$$\nu = \frac{1}{M}\sum_i \nu_{ii}^N = \frac{\partial\langle\bar{N}_0\rangle}{\partial m} = -\frac{\phi_N\phi_R(1-\sigma_c^2\nu)}{(1-\sigma_c^2\nu)\sigma_c^2\chi+1} - \frac{\phi_N(1-\phi_R)}{\sigma_c^2\chi} \tag{5.22}$$

$$\chi = \frac{1}{M}\sum_\alpha \chi_{\alpha\alpha}^R = \frac{\partial\langle\bar{R}_0\rangle}{\partial K} = \frac{\phi_N\phi_R\sigma_c^2\chi}{(1-\sigma_c^2\nu)\sigma_c^2\chi+1} + \frac{(1-\phi_N)\phi_R}{1-\sigma_c^2\nu} \tag{5.23}$$

5.3.1 With resource depletion

Two solutions are found by solving eq. (5.22) and eq. (5.23):

$$\phi_R - \phi_N = 0, \quad \chi = 0, \quad \nu = \frac{1}{\sigma_c^2 - 1} \tag{5.24}$$

$$\begin{aligned}&\phi_R - \phi_N > 0, \quad \chi = \phi_R - \phi_N, \\ &\nu = \frac{1 - 2\phi_N\sigma_c^2 + \phi_R\sigma_c^2 - \sqrt{1 + 2(1-2\phi_N)\phi_R\sigma_c^2 + \phi_R^2\sigma_c^4}}{2\sigma_c^4(\phi_R - \phi_N)}.\end{aligned} \tag{5.25}$$

5.3.2 Without resource depletion

In this case, the resource never vanishes so that we can fix $\phi_R = 1$ and solve eq. (5.22) and eq. (5.23). Two solutions are found:

$$1 - \phi_N = 0, \quad \chi = 0, \quad \nu = \frac{1}{\sigma_c^2 - 1} \tag{5.26}$$

$$1 - \phi_N > 0, \quad \chi = 1 - \phi_N, \quad \nu = \frac{1 - 2\phi_N\sigma_c^2 + \sigma_c^2 - \sqrt{1 + 2\sigma_c^2 - 4\phi_N\sigma_c^2 + \sigma_c^4}}{2\sigma_c^4(-1+\phi_N)}. \tag{5.27}$$

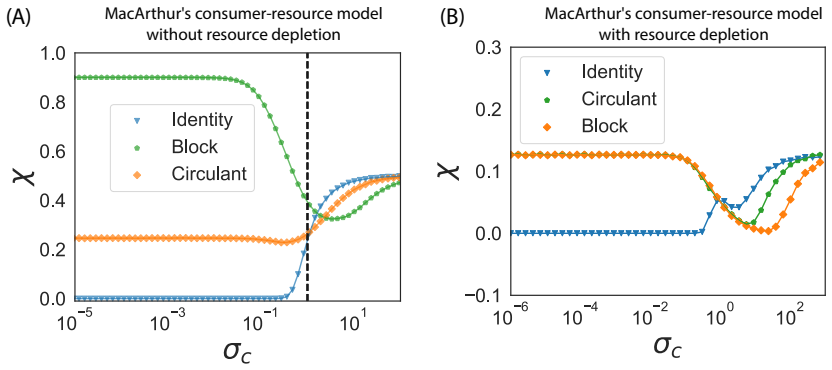

FIGURE 5.5: Comparison between numerical simulations(scatter points) and cavity solutions(solid lines) for χ at different σ_c for different cases. **(A)** CRM without resource depletion, eqs. (5.2). **(B)** CRM with resource depletion, eqs. (5.1). Note S^* and M^* are obtained from the numerical simulations, although in principle they could be obtained by solving the cavity equations directly.

Above two solutions are continuous at the transition point: $\chi = 0$ i.e. $\phi_N = 1$. Assume there is a small perturbation near the transition: $\phi_N = 1 - \epsilon$ and $\epsilon \ll 1$ and ν in eq. (5.27) can be expanded around ϵ. It is easy to check the ν in eq. (5.27) has the same expression as eq. (5.26) at the first order of ϵ. Therefore, only one solution exists:

$$\chi = 1 - \phi_N, \quad \nu = \frac{1 - 2\phi_N\sigma_c^2 + \sigma_c^2 - \sqrt{1 + 2\sigma_c^2 - 4\phi_N\sigma_c^2 + \sigma_c^4}}{2\sigma_c^4(-1 + \phi_N)} \tag{5.28}$$

The comparison between cavity solutions and numerical simulations for χ and ν are given in Figure 5.5 and Figure 5.4 respectively.

5.3.3 Without resource depletion and species extinction

In this case, both the resource and the species never vanish so that we can fix $\phi_R = 1$ and $\phi_N = 1$. Solving eq. (5.22) and eq. (5.23), only one solution is found:

$$\chi = 0, \quad \nu = \frac{1}{\sigma_c^2 - 1}. \tag{5.29}$$

5.3.4 Behavior in Three Regimes

To understand these solutions and behaviors better, it is helpful to consider three regimes: *Regime A* where $\chi = \phi_R - \phi_N = 0$, *Regime B* where χ becomes nonzero and species start to extinct, and *Regime C* where $\sigma_c \gg 1$ and it becomes a random ecosystem.

In *Regime B*, resource depletion has a significant effect on the system's feasibility, shown in Figure 5.4. With resource depletion, equation (5.25) shows there is a sudden change for the linear response function ν from *Regime A*: $\chi = 0$ to *Regime B* $\chi \neq 0$. As $\nu \sim \frac{1}{\phi_R - \phi_N}$, even a slightly decrease of the number of surviving species will induce a huge perturbation to the ecosystem, corresponding to a phase transition between *Regime A* and *Regime B* at $\sigma_c^* \sim 0.2$.

Without resource depletion, equation (5.28) shows the linear response function ν is continuous from *Regime A* to *Regime B*. There is a crossover instead of a phase transition there. The peak for the crossover is a finite value and can be calculated by taking the derivative of equation (5.28) over σ_c, ignoring the correlation between σ_c and ϕ_N. It happens approximately at $\sigma_c^* = \sqrt{4\phi_N - 2} \sim 1.04$, where $\phi_N = 0.77$ can be obtained from numerical simulation. The explanation for the difference from random matrix theory are provided in the main text and also the spectrums in Figure 5.3 and Figure 5.4.

Without resource and species depletion, as shown in equation (5.29), ν diverges at $\sigma_c^* = 1$, corresponding to $\lambda_{\min}$ reaching exactly zero. This result is also consistent with equation (5.44), predicted by random matrix theory, which ignores the effect of row or column deletions in the interaction matrix. This tells there do not exists any feasible solutions for the coexistence of M species and M resources. Therefore species must go extinct before $\sigma_c^* = 1$.

In *Regime C*, further increasing of σ_c after $\sigma_c > 1$, the σ_c^4 term in the square root becomes dominating and the the susceptibility ν behaves like a random ecosystem quickly, which explains the dramatic drop of the species packing shown in Figure 5.1. It indicates the ecosystem tends to a self-organized random state.

5.3.5 Solutions in Regime A and C

In *Regime A*($\sigma_c \ll 1$), for eqs. (5.1) with resource depletion, the solutions for the steady-states become,

$$R_0 = \max\left[0, m - z_N\right], \quad N_0 = \max\left[0, K + z_R\right]. \tag{5.30}$$

For eqs. (5.2) without resource depletion, the solutions for the steady-states become,

$$R_0 = m - z_N, \quad N_0 = \max\left[0, K + z_R\right]. \tag{5.31}$$

For ecosystems without resource and species extinction, the solutions for the steady-states become,

$$R_0 = m - z_N, \quad N_0 = K + z_R. \tag{5.32}$$

For *Regime C* ($\sigma_c \gg 1$), for eqs. (5.1) with resource depletion, the solutions for the steady-states become,

$$R_0 = \max\left[0, \frac{K - \mu\langle N\rangle + z_R}{1-\sigma_c^2\nu}\right], \quad N_0 = \max\left[0, \frac{\mu\langle R\rangle - m + z_N}{\sigma_c^2\chi}\right], \tag{5.33}$$

in agreement with the equations obtained in [ABM18b] for purely random interactions. For equations. (5.2) without resource depletion, the solutions for the steady-states become,

$$R_0 = \frac{K - \mu\langle N\rangle + z_R}{1-\sigma_c^2\nu}, \quad N_0 = \max\left[0, \frac{\mu\langle R\rangle - m + z_N}{\sigma_c^2\chi}\right]. \tag{5.34}$$

For ecosystems without resource and species extinction, the solutions for the steady-states become,

$$R_0 = \frac{K - \mu\langle N\rangle + z_R}{1-\sigma_c^2\nu}, \quad N_0 = \frac{\mu\langle R\rangle - m + z_N}{\sigma_c^2\chi}. \tag{5.35}$$

5.4 Correspondence between RMT and cavity solution

Our numerical simulations show that after the transition, our ecosystems are well described by purely random interactions. This suggests that we should be able to derive our cavity results using Random Matrix Theory (RMT). We now show that this is indeed the case. Our starting point are the average susceptibilities which are defined as:

$$\chi = \frac{1}{M}\sum_{\alpha\in\mathbf{M}} \chi^R_{\alpha\alpha} = \frac{1}{M}\sum_{\alpha\in\mathbf{M}^*} \chi^R_{\alpha\alpha} \tag{5.36}$$

$$\nu = \frac{1}{S}\sum_{i\in\mathbf{S}} \nu^N_{ii} = \frac{1}{S}\sum_{i\in\mathbf{S}^*} \nu^N_{ii}. \tag{5.37}$$

From the cavity calculations, we only care about $\chi^R_{\alpha\beta}$ and ν^N_{ij}, because the other susceptibilities are lower order in $1/M$.

We can combine these equations with (4.56) and (4.57) to obtain

$$\begin{aligned}\chi &= \frac{1}{M}\sum_{\alpha\in\mathbf{M}^*}\chi^R_{\alpha\alpha} = \frac{1}{M}\mathrm{Tr}(\chi^R_{\alpha\beta}) \qquad (5.38)\\ &= \frac{1}{M}\mathrm{Tr}(\delta_{\alpha\beta}) - \frac{1}{M}\mathrm{Tr}\left(\sum_{i\in\mathbf{S}^*}\sum_{j\in\mathbf{S}^*}\bar{C}^T_{\alpha i}A^{-1}_{ij}\bar{C}_{j\beta}\right)\\ &= \frac{M^*}{M} - \frac{1}{M}\mathrm{Tr}\left(\sum_{i\in\mathbf{S}^*}\sum_{j\in\mathbf{S}^*}A^{-1}_{ij}\bar{C}_{j\beta}\bar{C}^T_{\beta h}\right)\\ &= \frac{M^*}{M} - \frac{S^*}{M} = \phi_R - \gamma^{-1}\phi_N \qquad (5.39)\end{aligned}$$

We now show that the cavity solutions are consistent with results from RMT using equations (4.56) and(4.57) in Regime A and Regime C described in the main text.

5.4.1 Regime A: $\bar{\mathbf{C}} = \mathbb{1}$

This regime happens when $\sigma_c \ll 1$. Substituting, $\bar{\mathbf{C}} = \mathbb{1}$ into equations (4.56) and (4.57) yields

$$\chi = 0, \quad \nu = -1. \qquad (5.40)$$

This is consistent with the cavity solution equation (5.24) with $\sigma_c = 0$ since in this case $S^* = S = M$.

5.4.2 Regime C: $\bar{C}_{i\alpha}$ *i.i.d.* $\mathcal{N}(0, \sigma_c/\sqrt{M})$

In this regime, $\sigma_c \gg 1$. In this case, $A_{ij} = \sum_{\alpha\in\mathbf{S}^*}\bar{C}_{i\alpha}\bar{C}^T_{\alpha j}$ takes the form of a Wishart Matrix. We will exploit this to calculate χ and ν. Notice,

$$\nu = \frac{1}{S}\sum_{i\in\mathbf{S}^*}\nu^N_{ii} = -\frac{1}{S}\mathrm{Tr}(A^{-1}_{ij}) = -\frac{1}{S}\sum_{i=1}^{S^*}\lambda_i^{-1} \qquad (5.41)$$

where λ_i is the eigenvalue of A_{ij}. From the Marchenko-Pastur law [MP67b], we know that the eigenvalues of a random Wishart matrix obey the Marchenko-Pastur distribution. Substituting equation (5.6) into the expression for ν and replacing the sum with

an integral yields:

$$\begin{aligned} \nu &= -\frac{S^*}{S}\int_a^b \frac{1}{x}\rho(x)dx \\ &= -\frac{S^*}{S}\frac{a+b-2\sqrt{ab}}{4\sigma_c^2 y\sqrt{ab}} \\ &= -\frac{1}{\sigma_c^2}\frac{\phi_N}{\phi_R-\gamma^{-1}\phi_N} \end{aligned} \tag{5.42}$$

The second line of equation (5.42) is obtained by transferring the integral function to a complex analytic function and applying the residue theorem. This result is the same as the cavity solution equation (5.25) when $\sigma_c \gg 1$.

5.4.3 Regime B using the Stieltjes transformation

In Regime B, it hard to estimate the minimum eigenvalue. We can use Stieltjes transformation of information-plus-noise-type matrices which are well studied in wireless communications[DS07, CD11, LV+11], where $\mathbf{B}$ represents the information encoded in the signal and $\mathbf{C}$ is the noise in wireless communications. In this case, we have

$$\bar{C}_{i\alpha} = \mathbb{1} + C_{i\alpha}, \quad C_{i\alpha} \ i.i.d. \ \mathcal{N}(0, \sigma_c/\sqrt{M}).$$

$$A_{ij} = \sum_{\alpha\in M^*} \bar{C}_{i\alpha}\bar{C}^T_{\alpha j} = \sum_{\alpha\in M^*} C_{i\alpha}C^T_{\alpha j} + C_{i\alpha} + C^T_{\alpha i} + \mathbb{1} \tag{5.43}$$

Using **Theorem 1.1** in [DS07][DS07], the Stieltjes transform $m(z)$ of A_{ij} satisfies

$$\sigma_c^4 z m^3 - 2\sigma_c^2 z m + (\sigma_c^2 + z - 1)m - 1 = 0 \tag{5.44}$$

The asymptotic spectrum of A_{ij} can be obtained by $m(z)$, the solution of equation (5.44) with

$$\rho(x) = \lim_{\varepsilon\to 0^+} \frac{m(x-i\varepsilon) - m(x+i\varepsilon)}{2i\pi} \tag{5.45}$$

The result is shown in Figure 5.6. The minimum eigenvalue reaches 0 nearly at $\sigma_c^* = 1$, as predicted by the cavity solution.

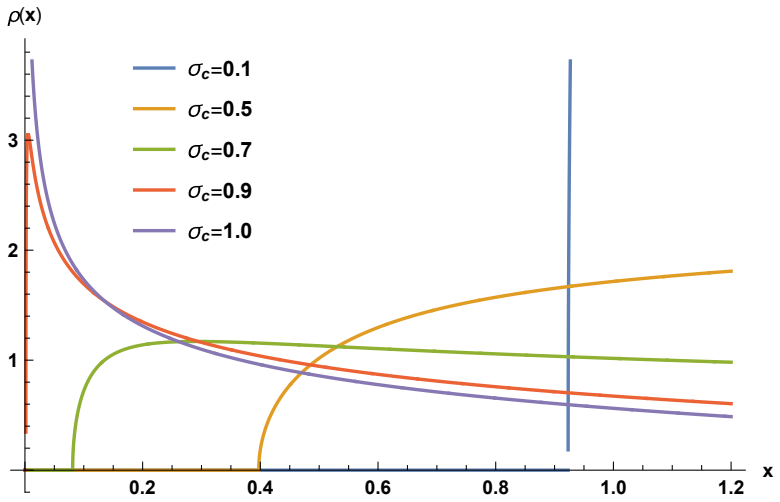

FIGURE 5.6: The asymptotic spectrum of A_{ij} for different values of σ_c by solving equation (5.45) numerically.

5.5 Summary

It is common practice in theoretical ecology to model ecosystems using random matrices. Yet it remains unclear if and when we can treat real communities as random ecosystems. Here, we investigated this question by generalizing May's analysis to consumer resource models and asking when the macroscopic, community level properties can be accurately predicted using random parameters. We found that introducing even modest amount of stochasticity into consumer preferences ensures that the macroscopic properties of diverse ecosystems will be indistinguishable from those of a completely random ecosystem.

We confirmed our analytic calculations using numerical simulations on CRMs with different types of resource dynamics and different classes of non-specific interactions. We also showed that despite the fact that random ecosystems can make accurate predictions about macroscopic properties like the average diversity or productivity, they will in general fail to capture species level details. This phenomena is well understood in the context of statistical physics where it is possible to predict thermodynamic quantities such as pressure and temperature even though one cannot accurately predict microstates.

These observations may help explain the surprising success of consumer resource models with random parameters in predicting the behavior of microbial ecosystems in the lab and natural environments [GLB+18, MCM20]. They also suggest that maybe possible to predict macroscopic ecosystem level properties like diversity or total biomass even when ecosystems are poorly characterized or have lots of missing data.

The foregoing analysis has several other interesting implications. First, it suggests that bottom-up engineering of complex ecosystems may be very difficult. As the number of components increases, small uncertainties in each of the interaction parameters may eventually overwhelm the designed interactions, and destabilize the intended steady state. Instead, such system are much more likely end up in a typical state which our theory suggests is much more stable than the intended designed state as ecosystems become more diverse.

Our work also suggests that in ecosystems well described by consumer resource models, crossing the May transition generically gives rise to typical random ecosystems rather than a marginal stable phase as was found in a recent analysis of the Generalized Lotka-Volterra model [BBC18b]. For this reason, even when the cumulative parameter uncertainties preclude a priori prediction of the detailed structure of the new state, methods from statistical physics and Random Matrix Theory can be employed to predict system-level properties [BABL18, SCG$^+$18]. For these reasons, we feel that further development of these methods are likely to play an important role in enabling top-down control of ecosystems and may help to identify assembly rules for microbial communities with many species [FHG17]

In this Letter, we only consider white noise, which is independently and identically added to all interaction components. In the future, it will be interesting to ask how other specialized noises, resulting from demographic stochasticity, phenotypic variation, can affect our results. Based on our experience, we expect that, even in these more complicated ecosystems, our conclusion will hold quite in the thermodynamics limit generically. But much more work needs to be done to confirm if this is really the case.

Chapter 6

Effects of Resource Dynamics

Few works recognize the importance of resource dynamics in shaping ecosystems. In theoretical ecology, often, it is assumed that all resources are the same. In contrast, we show that it is really important to also think about the dynamics of the resources if we really want to understand how much biodiversity an ecosystem can support. The linear resource dynamics we consider here are especially important in the realm of microbial ecosystems (Microbiomes). Understanding why the microbiomes we observe as so diverse is a fundamental question in biology. Our work can help design experiments to systematically understand how resource dynamics affects species coexistence patterns, and we are setting up collaborations to try to understand this better.

6.1 Model

Here we consider General consumer resource models (CRMs) describing the ecological dynamics of S species of consumers N_i ($i = 1, 2, \dots S$) that can consume M distinct resources R_α ($\alpha = 1, 2, \dots, M$). The rate at which species N_i consumes and depletes resource R_β is encoded in a matrix of consumer preferences $C_{i\beta}$. In order to survive, species have a minimum maintenance cost m_i. Equivalently, m_i can also be thought of as the death rate of species i in the absence of resources. These dynamics can be described using a coupled set of $M + S$ ordinary differential equations of the form

$$\begin{cases} \frac{dN_i}{dt} = N_i \sum_\beta C_{i\beta} R_\beta - N_i m_i \\ \\ \frac{dR_\alpha}{dt} = h_\alpha(R_\alpha) - \sum_j N_j C_{j\alpha} R_\alpha, \end{cases} \tag{6.1}$$

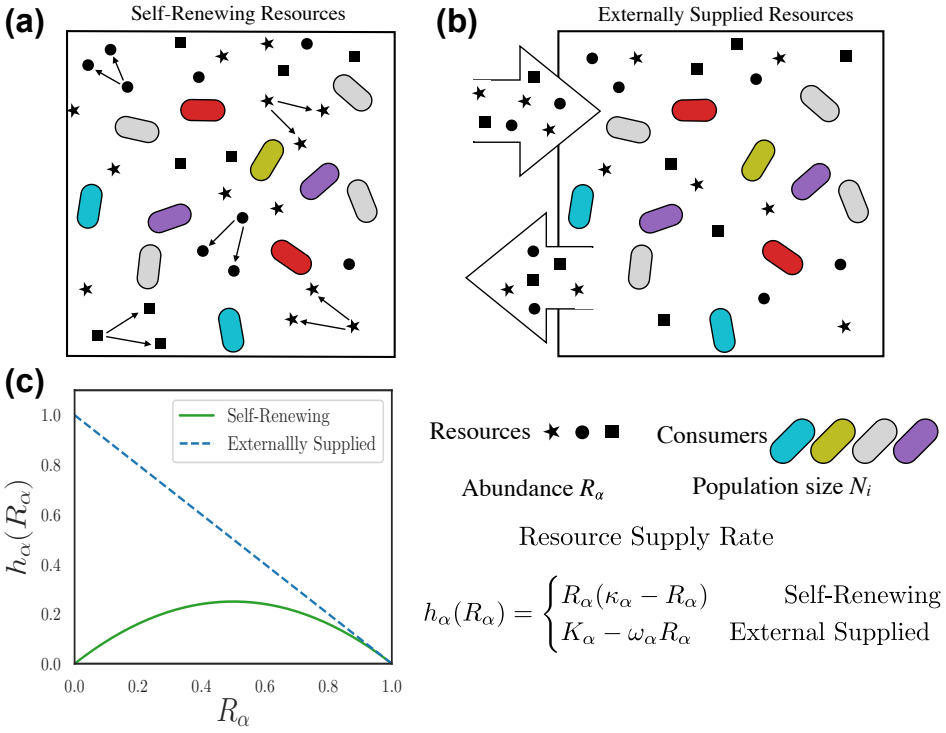

FIGURE 6.1: Schematic description for two types of resources. (a) Self-renewing resources (e.g. plants), which are replenished through organic reproduction; (b) Externally supplied resources (e.g. nutrients that sustain gut microbiota), which are replenished by a constant flux from some external source, and diluted at a constant rate; (c) The supply rate as a function of resource abundance for both choices, with $\kappa = \omega_\alpha = K_\alpha = 1$.

where $h_\alpha(R_\alpha)$ a function that describes the dynamics of the resources in the absence of any consumers (see Fig. 6.1).

For self-renewing resources (e.g. plants, animals), the dynamics can be described using logistic growth of the form

$$h_\alpha(R_\alpha) = R_\alpha(\kappa_\alpha - R_\alpha), \tag{6.2}$$

with κ the carrying capacity. While such resource dynamics is reasonable for biotic resources, abiotic resources such as minerals and small molecules cannot self-replicate and are usually supplied externally to the ecosystem (Fig. 6.1(b)). A common way to model this scenario is by using linearized resource dynamics of the form

$$h_\alpha(R_\alpha) = K_\alpha - \omega_\alpha R_\alpha. \tag{6.3}$$

Fig. 6.1(c) shows a plot of these two choices. Notice that the two resource dynamics behave very differently at low resource levels. The self-renewing resources can go extinct

and eventually disappear from the ecosystem while this is not true of externally supplied resources.

Recent research has shown some unexpected and interesting non-generic phenomena can appear in GCRMs in the presence of additional constraints on parameter values. A common choice of such constraints is the imposition of a "metabolic budget" on the consumer preference matrix [PTW17, LLL+19] tying the maintenance cost m_i to the total consumption capacity $\sum_\beta C_{i\beta}$ [TM17, AF19]. These metabolic tradeoffs can be readily incorporated into the cavity calculations and have significant impacts on species packing as will be discussed below.

6.2 Cavity solution

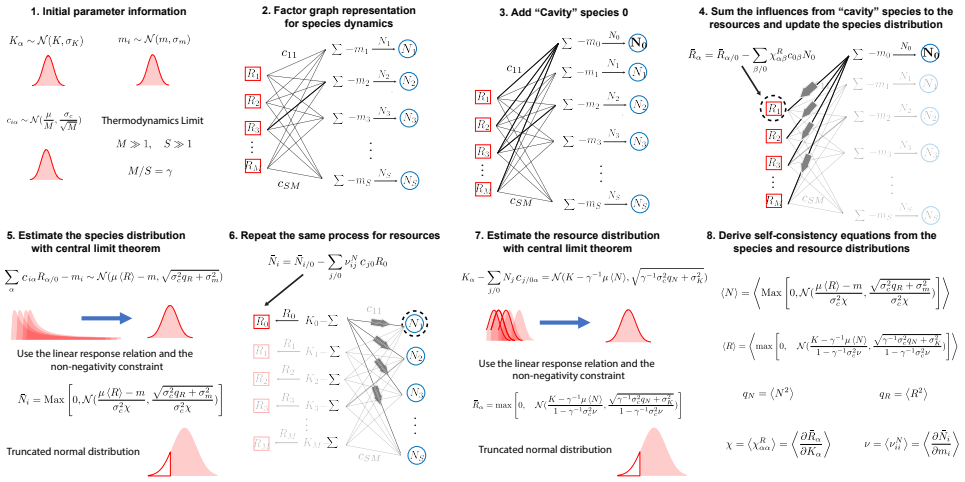

FIGURE 6.2: Schematic outlining steps in cavity solution. **1.** The initial parameter information consists of the probability distributions for the mechanistic parameters: K_α, m_i and $C_{i\alpha}$. We assume they can be described by their first and second moments. **2.** The species dynamics $N_i(\sum_\alpha C_{i\alpha}R_\alpha - m_i)$ in eqs. (6.4) are expressed as a factor graph. **3.** Add the "Cavity" species 0 as the perturbation. **4.** Sum the resource abundance perturbations from the "Cavity" species 0 at steady state and update the species abundance distribution to reflect the new steady state. **5.** Employing the central limit theorem, the backreaction contribution from the "cavity" species 0 and the non-negativity constraint, the species distribution is expressed as a truncated normal distribution. **6.** Repeat **Step 2-4** for the resources. **7.** The resource distribution is the ratio distribution from the ratio of two normal variables K_α and $\omega_\alpha + \sum_i N_i C_{i\alpha}$. **8.** The self-consistency equations are obtained from the species and resource distributions. Note that $\gamma^{-1}\sigma_c^2\nu\langle R\rangle$ in the dominator of $\langle R\rangle$ is from the correlation between N_i and $C_{i\alpha}$ in $\sum_i N_i C_{i\alpha}$.

6.2.1 Model setup

In this section, we derive the cavity solution to the linear resource dynamics (eq. (6.1).

$$\begin{cases} \frac{dN_i}{dt} = N_i \left(\sum_\beta C_{i\beta} R_\beta - m_i \right) \\ \\ \frac{dR_\alpha}{dt} = K_\alpha - \omega_\alpha R_\alpha - \sum_j N_j C_{j\alpha} R_\alpha \end{cases} \tag{6.4}$$

Note that here we follow closely our derivation in [ABM18b, MCWMI19]. The main difference is that here we consider linear resource dynamics, which as we will see below, makes the problem much more technically challenging.

Consumer preference $C_{i\alpha}$ are random variables drawn from a Gaussian distribution with mean μ/M and variance σ_c^2/M. They can be deposed into $C_{i\alpha} = \mu/M + \sigma_c d_{i\alpha}$, where the fluctuating part $d_{i\alpha}$ obeys

$$\langle d_{i\alpha} \rangle = 0 \tag{6.5}$$

$$\langle d_{i\alpha} d_{j\beta} \rangle = \frac{\delta_{ij}\delta_{\alpha\beta}}{M}. \tag{6.6}$$

We also assume that both the carrying capacity K_α and the minimum maintenance cost m_i are independent Gaussian random variables with mean and covariance given by

$$\langle K_\alpha \rangle = K \tag{6.7}$$

$$\mathrm{Cov}(K_\alpha, K_\beta) = \delta_{\alpha\beta}\sigma_K^2 \tag{6.8}$$

$$\langle m_i \rangle = m \tag{6.9}$$

$$\mathrm{Cov}(m_i, m_j) = \delta_{ij}\sigma_m^2 \tag{6.10}$$

Let $\langle R \rangle = \frac{1}{M}\sum_\beta R_\beta$ and $\langle N \rangle = \frac{1}{S}\sum_j N_j$ be the average resource and average species abundance, respectively. With all these defined, we can re-write eqs. (6.4) as

$$\frac{dN_i}{dt} = N_i \left\{ \mu \langle R \rangle - m + \sum_\beta \sigma_c d_{i\beta} R_\beta - \delta m_i \right\} \tag{6.11}$$

$$\frac{dR_\alpha}{dt} = K + \delta K_\alpha - \left[\omega_\alpha + \gamma^{-1}\mu \langle N \rangle + \sum_j \sigma_c d_{j\alpha} N_j \right] R_\alpha \tag{6.12}$$

where $\delta K_\alpha = K_\alpha - K, \delta m_i = m_i - m$ and $\gamma = M/S$. As noted in the main text, the basic idea of cavity method is to relate an ecosystem with $M+1$ resources (variables) and

$S+1$ species (inequality constraints) to that with M resources and S species. Following eq. (6.11) and eq. (6.12), one can write down the ecological model for the $(M+1, S+1)$ system where resource R_0 and species N_0 are introduced to the (M, S) system as:

$$\frac{dN_0}{dt} = N_0 \left\{ \mu \langle R \rangle - m + \sum_\beta \sigma_c d_{0\beta} R_\beta - \delta m_0 \right\} \tag{6.13}$$

$$\frac{dR_0}{dt} = K + \delta K_0 - \left[\omega_0 + \gamma^{-1} \mu \langle N \rangle + \sum_j \sigma_c d_{j0} N_j \right] R_0 \tag{6.14}$$

6.2.2 Perturbations in cavity solution

Following the same procedure as in [ABM18b], we introduce the following susceptibilities:

$$\chi^R_{\alpha\beta} = -\frac{\partial \bar{R}_\alpha}{\partial \omega_\beta} \tag{6.15}$$

$$\chi^N_{i\alpha} = -\frac{\partial \bar{N}_i}{\partial \omega_\alpha} \tag{6.16}$$

$$\nu^R_{\alpha i} = \frac{\partial \bar{R}_\alpha}{\partial m_i} \tag{6.17}$$

$$\nu^N_{ij} = \frac{\partial \bar{N}_i}{\partial m_j} \tag{6.18}$$

where we denote $\bar{X}$ as the steady-state value of X. Recall that the goal is to derive a set of self-consistency equations that relates the ecological system characterized by $M+1$ resources (variables) and $S+1$ species (constraints) to that with the new species and new resources removed: $(S+1, M+1) \to (S, M)$. To simplify notation, let $\bar{X}_{\setminus 0}$ denote the steady-state value of quantity X in the absence of the new resource and new species. Since the introduction of a new species and resource represents only a small (order $1/M$) perturbation to the original ecological system, we can express the steady-state species and resource abundances in the $(S+1, M+1)$ system with a first-order Taylor expansion around the (S, M) values. We note that the new terms $\sigma_c d_{i0} R_0$ in Eq. eq. (6.12) and $\sigma_c d_{0\alpha} N_0$ in eq. (6.11) can be treated as perturbations to m_i, and K_α, respectively, yielding:

$$\bar{N}_i = \bar{N}_{i/0} - \sigma_c \sum_{\beta/0} \chi^N_{i\beta} d_{0\beta} \bar{N}_0 - \sigma_c \sum_{j/0} \nu^N_{ij} d_{j0} \bar{R}_0 \tag{6.19}$$

$$\bar{R}_\alpha = \bar{R}_{\alpha/0} - \sigma_c \sum_{\beta/0} \chi^R_{\alpha\beta} d_{0\beta} \bar{N}_0 - \sigma_c \sum_{j/0} \nu^R_{\alpha j} d_{j0} \bar{R}_0 \tag{6.20}$$

Note $\sum_{j/0}$ and $\sum_{\beta/0}$ mean the sum excludes the new species 0 and the new resource 0. The next step is to plug eq. (6.19) and eq. (6.20) into eq. (6.13) and eq. (6.14) and solve for the steady-state value of N_0 and R_0.

6.2.3 Self-consistency equations for species

For the new cavity species, the steady equation takes the form

$$\begin{aligned} 0 = \bar{N}_0[\mu \langle R \rangle - m - \sigma_c^2 \bar{N}_0 \sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} d_{0\alpha} d_{0\beta} - \sigma_c^2 \bar{R}_0 \sum_{\beta/0,j/0} \nu^R_{\beta j} d_{0\beta} d_{0j} \\ + \sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} + \sigma_c d_{00} \bar{R}_0 - \delta m_0] \end{aligned} \tag{6.21}$$

Notice that each of the sums in this equation is the sum over a large number of weak correlated random variables, and can therefore be well approximated by Gaussian random variables for large enough M and S. We can calculate the sum of the random variables:

$$\sum_{\beta/0,j/0} \nu^R_{\beta j} d_{0\beta} d_{0j} = \frac{1}{M} \sum_{\beta/0,j/0} \nu^R_{\beta j} \delta_{j0} \delta_{\beta 0} = 0 \tag{6.22}$$

$$\sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} d_{0\alpha} d_{0\beta} = \frac{1}{M} \sum_{\alpha/0,\beta/0} \chi^R_{\alpha\beta} \delta_{\alpha\beta} = \frac{1}{M} \sum_{\alpha} \chi^R_{\alpha\alpha} = \frac{1}{M} \mathrm{Tr}(\chi^R_{\alpha\beta}) = \chi \tag{6.23}$$

where χ is the average susceptibility. Using these observations about above sums, we obtain

$$0 = \bar{N}_0 \left[\mu \langle R \rangle - m - \sigma_c^2 \chi \bar{N}_0 + \sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} + \sigma_c d_{00} \bar{R}_0 - \delta m_0 \right] + \mathcal{O}(M^{-1/2}), \tag{6.24}$$

Employing the Central Limit Theorem, we introduce an auxiliary Gaussian variable z_N with zero mean and unit variance and rewrite this as

$$\sum_{\beta/0} \sigma_c d_{0\beta} \bar{R}_{\beta/0} + \sigma_c d_{0\beta} \bar{R}_0 - \delta m_0 = z_N \sqrt{\sigma_c^2 q_R + \sigma_m^2}, \tag{6.25}$$

where q_R is the second moment of the resource distribution,

$$q_R = \frac{1}{M} \sum_{\beta} R_\beta^2.$$

We can solve eq. (6.24) in terms of the quantities just defined:

$$\mu \langle R \rangle - m - \sigma_c^2 \chi \bar{N}_0 + \sqrt{\sigma_c^2 q_R + \sigma_m^2} z_N \leq 0 \tag{6.26}$$

Inverting this equation one gets the steady state of species

$$\bar{N}_0 = \max \left[0, \quad \frac{\mu \langle R \rangle - m + \sqrt{\sigma_c^2 q_R + \sigma_m^2} z_N}{\sigma_c^2 \chi} \right] \tag{6.27}$$

which is a truncated Gaussian.

Let $y = \max\left(0, \frac{a}{b} + \frac{c}{b} z\right)$, with z being a Gaussian random variable with zero mean and unit variance. Then its j-th moment is given by

$$\begin{aligned} \langle y^j \rangle &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(\frac{c}{b} x + \frac{a}{b} \right)^j dx & (6.28) \\ &= \left(\frac{c}{b} \right)^j \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(x + \frac{a}{c} \right)^j dx & (6.29) \\ &= \left(\frac{c}{b} \right)^j w_j(\frac{a}{c}) & (6.30) \end{aligned}$$

here we define $w_j(\frac{a}{c}) = \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{c}}^{\infty} e^{-\frac{x^2}{2}} \left(x + \frac{a}{c} \right)^j dx$

With this we can easily write down the self-consistency equations for the fraction of non-zero species and resources as well as the moments of their abundances at the steady state:

$$\begin{aligned} \phi_N &= \frac{S^*}{S} = w_0 \left(\frac{\mu \langle R \rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}} \right) & (6.31) \\ \langle N \rangle &= \frac{1}{S} \sum_j N_j = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi} \right) w_1(\frac{\mu \langle R \rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}) & (6.32) \\ q_N &= \frac{1}{S} \sum_j N_j^2 = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi} \right)^2 w_2(\frac{\mu \langle R \rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}) & (6.33) \end{aligned}$$

Note that S^* is the number of surviving species at the steady state.

6.2.4 Self-consistency equations for resources

We now derive the equations for the steady-state of the resource dynamics. Inserting eq. (6.20) into eq. (6.14) gives:

$$\begin{aligned} 0 =& K + \delta K_0 - \bar{R}_0[\omega + \gamma^{-1}\mu \langle N \rangle - \sigma_c^2 \bar{N}_0 \sum_{\beta/0,j/0} \chi^N_{j\beta} d_{j0} d_{0\beta} \\ & - \sigma_c^2 \bar{R}_0 \sum_{i/0,j/0} \nu^N_{ij} d_{0i} d_{0j} + \sum_{j/0} \sigma_c d_{j0} \bar{N}_{j/0} + \sigma_c d_{00} \bar{N}_0 + \delta\omega_0] \end{aligned} \tag{6.34}$$

We can simplify the sums by averaging over the random variables:

$$\sum_{\beta/0,j/0} \chi^N_{j\beta} d_{j0} d_{0\beta} = \frac{1}{M} \sum_{\beta/0,j/0} \chi^N_{j\beta} \delta_{j0} \delta_{\beta 0} = 0 \tag{6.35}$$

$$\sum_{i/0,j/0} \nu^N_{ij} d_{0i} d_{0j} = \frac{1}{M} \sum_{i/0,j/0} \nu^N_{ij} \delta_{ij} = \frac{1}{M} \sum_i \nu^N_{ii} = \frac{1}{M} \mathrm{Tr}(\nu^N_{ij}) = \gamma^{-1}\nu \tag{6.36}$$

where ν is the average susceptibility. Finally, note that we can write

$$\delta\omega_0 + \sum_j \sigma_c d_{j0} N_j = z_R \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2}, \tag{6.37}$$

where we have introduced another auxiliary Gaussian variable z_R with zero mean and unit variance and q_N is the second moment of the resource distribution defined in eq. (6.58), Using these observations, we obtain a quadratic expression for the resource.

$$K + \delta K_0 - (\omega_0 + \gamma^{-1}\mu \langle N \rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)\bar{R}_0 + \gamma^{-1}\sigma_c^2 \nu \bar{R}_0^2 = 0 \tag{6.38}$$

6.2.4.1 Cavity solution: without backreaction

As discussed in the main text, we cannot solve the full resource equations exactly. For this reason, we perform an expansion, as a start, we calculate this equation by setting $\nu = 0$ in the resource equation. This is equivalent in the TAP language of ignoring the backreaction term.

Under this assumption, the quadratic equation for the resource, simply becomes a linear equation that can be re-arranged to give

$$\bar{R}_\alpha = \frac{K + \delta K_\alpha}{\omega + \gamma^{-1}\mu \langle N \rangle + z_R \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2}} \tag{6.39}$$

Assuming the fluctuations in the denominator is small, *i.e.* $\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} \ll \omega + \gamma^{-1}\mu\langle N\rangle$, we can do a first-order Taylor expansion around the mean value and also ignore the coupling term between δK_α and z_R:

$$\bar{R}_\alpha = \frac{K + \delta K_\alpha}{\omega + \gamma^{-1}\mu\langle N\rangle} - \frac{K\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2}}{(\omega + \gamma^{-1}\mu\langle N\rangle)^2} z_R \tag{6.40}$$

With all these approximations, we get the first two moments of the steady-state resource abundance distribution:

$$\langle R\rangle = \frac{K}{\omega + \gamma^{-1}\mu\langle N\rangle} \tag{6.41}$$

$$q_R = \langle R\rangle^2 + \frac{\sigma_K^2}{(\omega + \gamma^{-1}\mu\langle N\rangle)^2} + \frac{K^2(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2)}{(\omega + \gamma^{-1}\mu\langle N\rangle)^4} \tag{6.42}$$

The susceptibility is given by:

$$\begin{aligned}\chi &= -\left\langle \frac{\partial \bar{R}_\alpha}{\partial w_\alpha} \right\rangle = \left\langle \frac{K_\alpha}{(\omega_\alpha + \sum_j c_{j\alpha}\bar{N}_j)^2} + \frac{2K\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2}}{(\omega + \gamma^{-1}\mu\langle N\rangle)^3} z_R \right\rangle \\ &= \frac{K}{(\omega + \gamma^{-1}\mu\langle N\rangle)^2}\end{aligned}$$

Combined with self-consistency equations for species, we get the full set of :

$$\phi_N = w_0\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right), \qquad \chi = \frac{K}{(\omega + \gamma^{-1}\mu\langle N\rangle)^2} \tag{6.43}$$

$$\langle N\rangle = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi}\right) w_1\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right), \qquad \langle R\rangle = \frac{K}{\omega + \gamma^{-1}\mu\langle N\rangle} \tag{6.44}$$

$$q_N = \left(\frac{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}{\sigma_c^2 \chi}\right)^2 w_2\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R + \sigma_m^2}}\right), \tag{6.45}$$

$$q_R = \langle R\rangle^2 + \frac{\sigma_K^2}{(\omega + \gamma^{-1}\mu\langle N\rangle)^2} + \frac{K^2(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2)}{(\omega + \gamma^{-1}\mu\langle N\rangle)^4}. \tag{6.46}$$

6.2.4.2 Cavity solution: with backreaction correction

We start again with the full resource equation:

$$K + \delta K_0 - (\omega_0 + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)\bar{R}_0 + \gamma^{-1}\sigma_c^2 \nu \bar{R}_0^2 = 0 \tag{6.47}$$

Since $R_0 > 0$ and $\nu < 0$, the solution of eq. (6.38) gives:

$$
\begin{aligned}
R_0 =& \frac{\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R}{2\gamma^{-1}\sigma_c^2\nu} \\
&- \frac{\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)^2 - 4\gamma^{-1}\nu\sigma_c^2(K + \delta K_0)}}{2\gamma^{-1}\sigma_c^2\nu}
\end{aligned} \tag{6.48}
$$

For the 1^{st} order expansion, we assume $4\gamma^{-1}\nu\sigma_c^2\delta K_0 + 2\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R + (\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2) z_R^2 \ll (\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K$ and do a 1st order expansion around the mean of the form:

$$
\begin{aligned}
& \sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)^2 - 4\gamma^{-1}\nu\sigma_c^2(K + \delta K_0)} \\
=\ & \sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K} \\
+\ & \frac{(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2) z_R^2 + 2(\omega + \gamma^{-1}\mu\langle N\rangle)\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R - 4\gamma^{-1}\nu\sigma_c^2\delta K_0}{2\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K}}
\end{aligned}
$$

Using these expressions, the moments of their abundances at steady state can be calculated yielding:

$$
\begin{aligned}
\langle R\rangle =& \frac{\omega + \gamma^{-1}\mu\langle N\rangle}{2\gamma^{-1}\sigma_c^2\nu} - \frac{\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K}}{2\gamma^{-1}\sigma_c^2\nu} \\
&- \frac{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2}{4\gamma^{-1}\sigma_c^2\nu\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K}}
\end{aligned} \tag{6.49}
$$

$$
\begin{aligned}
q_R =\ & \langle R\rangle^2 + \frac{(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2)^2 + 8(\gamma^{-1}\nu\sigma_c^2\sigma_K)^2}{2(2\gamma^{-1}\sigma_c^2\nu)^2[(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K]} \\
+\ & \frac{(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2)[\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K} - (\omega + \gamma^{-1}\mu\langle N\rangle)]^2}{(2\gamma^{-1}\sigma_c^2\nu)^2[(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K]}
\end{aligned} \tag{6.50}
$$

From eq. (6.48),

$$
\frac{\partial R_0}{\partial\omega} = \frac{1}{2\gamma^{-1}\sigma_c^2\nu}\left\{1 - \frac{\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R}{\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)^2 - 4\gamma^{-1}\nu\sigma_c^2(K + \delta K_0)}}\right\} \tag{6.51}
$$

The term inside the bracket can be expanded as:

$$
\begin{aligned}
& \frac{\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R}{\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)^2 - 4\gamma^{-1}\nu\sigma_c^2(K + \delta K_0)}} \\
\approx\ & \frac{\omega + \gamma^{-1}\mu\langle N\rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R}{\sqrt{(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K}}\left[1 - \frac{(\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2) z_R^2 + 2(\omega + \gamma^{-1}\mu\langle N\rangle)\sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R - 4\gamma^{-1}\nu\sigma_c^2\delta K_0}{2(\omega + \gamma^{-1}\mu\langle N\rangle)^2 - 4\gamma^{-1}\nu\sigma_c^2 K}\right]
\end{aligned} \tag{6.52}
$$

The susceptibilities are given by averaging eq. (6.51)

$$\chi = -\left\langle \frac{\partial R}{\partial \omega} \right\rangle \tag{6.53}$$

$$\equiv \frac{1}{2\gamma^{-1}\nu\sigma_c^2}\{1 - \frac{\omega+\gamma^{-1}\mu\langle N\rangle}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K}} + \frac{3(\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2)(\omega+\gamma^{-1}\mu\langle N\rangle)}{2[(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K]^{3/2}}\} \tag{6.54}$$

$$\nu = \left\langle \frac{\partial N}{\partial m} \right\rangle = -\frac{\phi_N}{\sigma_c^2\chi} \tag{6.55}$$

Combined with self-consistency equations for species, get the full set of 1^{st} order self-consistency equations:

$$\phi_N = w_0\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}\right) \tag{6.56}$$

$$\langle N\rangle = \left(\frac{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}{\sigma_c^2\chi}\right) w_1\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}\right) \tag{6.57}$$

$$q_N = \left(\frac{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}{\sigma_c^2\chi}\right)^2 w_2\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}\right) \tag{6.58}$$

$$\langle R\rangle = \frac{\omega+\gamma^{-1}\mu\langle N\rangle}{2\gamma^{-1}\sigma_c^2\nu} - \frac{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K}}{2\gamma^{-1}\sigma_c^2\nu} - \frac{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}{4\gamma^{-1}\sigma_c^2\nu\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K}} \tag{6.59}$$

$$q_R = \langle R\rangle^2 + \frac{(\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2)^2+8(\gamma^{-1}\nu\sigma_c^2\sigma_K)^2}{2(2\gamma^{-1}\sigma_c^2\nu)^2[(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K]} + \frac{(\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2)[\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K}-(\omega+\gamma^{-1}\mu\langle N\rangle)]^2}{(2\gamma^{-1}\sigma_c^2\nu)^2[(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K]} \tag{6.60}$$

$$\chi = -\frac{1}{2\gamma^{-1}\nu\sigma_c^2}\left\{1 - \frac{\omega+\gamma^{-1}\mu\langle N\rangle}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K}} + \frac{3(\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2)(\omega+\gamma^{-1}\mu\langle N\rangle)}{2[(\omega+\gamma^{-1}\mu\langle N\rangle)^2-4\gamma^{-1}\nu\sigma_c^2 K]^{3/2}}\right\} \tag{6.61}$$

$$\nu = -\frac{\phi_N}{\sigma_c^2\chi} \tag{6.62}$$

6.2.5 Comparison between with and without backreaction

We can reduce the cavity solution with backreaction to the simpler one when σ_c is large. In fact all the complexity of cavity solution with backreaction comes from the expression for eq. (6.48):

$$R_0 = \frac{\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R}{2\gamma^{-1}\sigma_c^2\nu} - \frac{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R)^2-4\gamma^{-1}\nu\sigma_c^2(K+\delta K_0)}}{2\gamma^{-1}\sigma_c^2\nu}$$

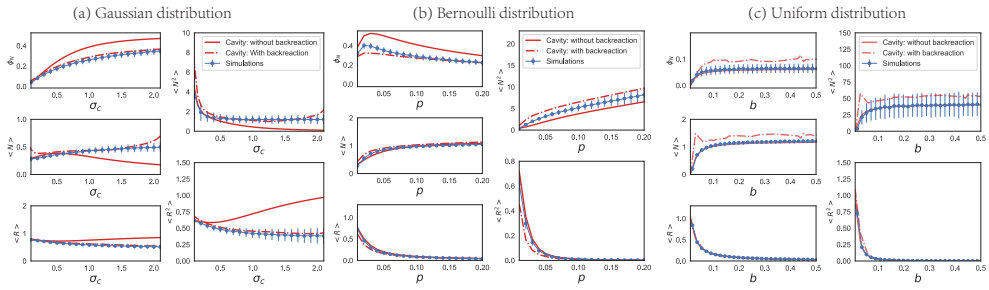

FIGURE 6.3: Comparison of numerics and cavity solutions with and without the back-reaction term as a function of σ_c. $\phi_N = \frac{S^*}{S}$ is the fraction of surviving species. $\langle N \rangle, \langle N^2 \rangle, \langle R \rangle$ and $\langle R^2 \rangle$ are the first and second moments of the species and resources distribution respectively. The simulations details can be found at the Appendix B.2. $\mathbf{C}$ is sampled either from a Gaussian, Bernoulli, or uniform distribution as indicated.

However, if we assume $(\omega + \gamma^{-1}\mu \langle N \rangle + \sqrt{\gamma^{-1}\sigma_c^2 q_N + \sigma_\omega^2} z_R)^2 \gg -4\gamma^{-1}\nu\sigma_c^2(K + \delta K_0)$, we can expand the second term following $\sqrt{1-x} \approx 1 - \frac{x}{2} - \frac{x^2}{8} + \mathcal{O}(x^3)$.

$$R_0 = \frac{K+\delta K_0}{\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2} z_R} + \frac{\gamma^{-1}\sigma_c^2\nu(K+\delta K_0)^2}{(\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2} z_R)^3} \tag{6.63}$$

The first term of above equation is the cavity solution without backreaction.

6.2.6 Comparing the cavity solutions to numerical simulations

We show a comparison between theoretical and numerical results for different choices of how to sample the consumption matrix in Fig. 6.5 and Fig. 6.3. These figures show that the cavity solution with backreaction performs better for the Gaussian and Bernoulli cases. However, in the uniform case, the cavity solution without backreaction matches with numerical simulations perfectly, while the cavity solution with backreaction performs worse than without backreaction. In the section 6.2.5, we have shown the cavity solution with backreaction can be reduced to the cavity solution without backreaction and hence should be a more robust solution. So why does it perform badly in the uniform case? The reason is that in the uniform case $\mu = Mb/2 \gg 1$ when the system size M is large, leading to $|\chi| \sim \frac{1}{(\omega+\gamma^{-1}\mu\langle N\rangle)^2} \ll 1$. From eqs. (6.57, 6.58), we see that both $\langle N \rangle$ and $\langle N^2 \rangle$ depends on $\frac{1}{\chi} \gg 1$ and the numerical solver becomes unstable.

6.3 An upper bound for species packing

By analyzing the susceptibilities in the full Cavity solutions, an upper bound for species packing can be derived for both resource dynamics in GCRMs. The derivations can also

be extended to the case where metabolic tradeoffs impose hard or soft constraints on the parameter values.

6.3.1 Externally supplied resource dynamics

The response functions χ and ν can be written as:

$$\chi = -\frac{1}{2\gamma^{-1}\sigma_c^2\nu}\left\{1-\left\langle \frac{\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R)^2-4\gamma^{-1}\nu\sigma_c^2(K+\delta K_0)}}\right\rangle\right\} \tag{6.64}$$

$$\nu = -\frac{\phi_N}{\sigma_c^2\chi} \tag{6.65}$$

Substituting eq. (6.65) into eq. (6.64) and rearranging yields

$$\gamma^{-1}\phi_N = \frac{1}{2}\left\{1-\left\langle \frac{\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle+\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_\omega^2}z_R)^2-4\gamma^{-1}\nu\sigma_c^2(K+\delta K_0)}}\right\rangle\right\}. \tag{6.66}$$

The numerator of the term in angle brackets is the total depletion rate for a given resource when it is first added to the system. Depletion rates are always positive in this model, so the right-hand side is always less than 1/2. Noticing $\gamma = \frac{M}{S}$, $\phi_N = S^*/S$, $\chi > 0$, we immediately obtain an upper bound on $\frac{S^*}{M}$:

$$\frac{1}{2} > \frac{S^*}{M}. \tag{6.67}$$

6.3.2 Self-renewing(MacArthur's) resource dynamics

Using the analytical expressions χ, ν and self-consistent equations in ref. [CMIM19], we can derive the following expressions:

$$\langle N\rangle = \left(\frac{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}{\sigma_c^2(\phi_R-\gamma^{-1}\phi_N)}\right) w_1\left(\frac{\mu\langle R\rangle - m}{\sqrt{\sigma_c^2 q_R+\sigma_m^2}}\right), \tag{6.68}$$

$$\langle R\rangle = \left(\frac{\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_K^2}}{\phi_R(\phi_R-\gamma^{-1}\phi_N)^{-1}}\right) w_1\left(\frac{\kappa-\gamma^{-1}\mu\langle N\rangle}{\sqrt{\gamma^{-1}\sigma_c^2 q_N+\sigma_K^2}}\right). \tag{6.69}$$

To derive bounds, we consider various limits of these expressions. First, consider the case were we put many species $S\to\infty$ into the ecosystem with fixed number of resources M, (i.e $\gamma = \frac{M}{S}\to 0$). In order to keep $\langle N\rangle$ positive, we must have $\phi_R - \gamma^{-1}\phi_N > 0$, giving an upper bound:

$$1 \geq \frac{M^*}{M} > \frac{S^*}{M} \tag{6.70}$$

6.3.3 Externally supplied resources with metabolic tradeoffs

Here we consider two kinds of constraints on the parameters, encoding metabolic tradeoffs. In the first, the maintenance cost $m_i = m$ is the same for all species, and the sum of the consumption preferences is constrained to equal some fixed "enzyme budget" E that is nearly the same for all species:

$$\sum_\alpha C_{i\alpha} = E + \delta E_i \tag{6.71}$$

where δE_i is a small random variable with mean zero and variance σ_E^2. A hard constraint can be generated by taking $\sigma_E = 0$.

The second kind of constraint does not make any assumptions about $C_{i\alpha}$, but assigns a cost $\tilde{m}$ to every unit of consumption capacity, so that

$$m_i = (1+\epsilon_i)\tilde{m}\sum_\alpha C_{i\alpha} + \delta m_i \tag{6.72}$$

where ϵ_i and δm_i are small random variables with mean zero and variances σ_ϵ^2 and σ_m^2, respectively. A hard constraint can be generated by taking $\sigma_\epsilon = \sigma_m = 0$.

In the simplest way of setting up the first constraint, the equilibrium equations actually reduce to the same form as the second. Specifically, one usually generates a consumer preference matrix satisfying the constraint by first generating an i.i.d. matrix $\tilde{C}_{i\alpha}$, and then setting $C_{i\alpha} = (E+\delta E_i)\tilde{C}_{i\alpha}/\sum_\beta \tilde{C}_{i\beta}$. The resulting dynamics can be written as:

$$\frac{dN_i}{dt} = N_i\left[\sum_\alpha (E+\delta E_i)\frac{\tilde{C}_{i\alpha}}{\sum_\beta \tilde{C}_{i\beta}}R_\alpha - m_i\right] \tag{6.73}$$

$$= \frac{N_i(E+\delta E_i)}{\sum_\beta \tilde{C}_{i\beta}}\left[\sum_\alpha \tilde{C}_{i\alpha}R_\alpha - m\frac{\sum_\beta \tilde{C}_{i\beta}}{E+\delta E_i}\right]. \tag{6.74}$$

Dropping the tilde's, we can write the equilibrium condition in the same form that results from the second kind of constraint:

$$0 = N_i\{\sum_\alpha C_{i\alpha}[R_\alpha - (1+\epsilon_i)\tilde{m}] - \delta m_i\} \tag{6.75}$$

with

$$\tilde{m} = \frac{m}{E} \tag{6.76}$$

$$\epsilon_i = -\frac{\delta E_i}{E} \tag{6.77}$$

$$\delta m_i = 0. \tag{6.78}$$

Inspection of Equation 6.75 immediately reveals an important novelty: now when we add a new resource as part of the cavity protocol, the perturbation to the growth rate can either be positive or negative, depending on the sign of $[R_\alpha - (1+\epsilon_i)\tilde{m}]$. This turns out to be the crucial factor that prevents the proof of the $S^*/M < 1/2$ bound from going through, regardless of the size of σ_ϵ or σ_m.

Following the same steps as above, we arrive at the following set of equilibrium conditions for the new species N_0 and resource R_0:

$$0 = \bar{N}_0\left[\mu\langle R\rangle - \mu\tilde{m} + \sigma_N z_N - \sigma_c^2\chi\bar{N}_0\right] \tag{6.79}$$

$$0 = K + \delta K_0 - (\omega + \gamma^{-1}\mu\langle N\rangle + \sigma_R z_R + \gamma^{-1}\sigma_c^2\nu\tilde{m})\bar{R}_0 + \gamma^{-1}\sigma_c^2\nu\bar{R}_0^2 \tag{6.80}$$

where

$$\sigma_N^2 = \sigma_m^2 + \sigma_c^2[q_R - 2\tilde{m}\langle R\rangle + \tilde{m}^2(1+\sigma_\epsilon^2)] \tag{6.81}$$

$$\sigma_R^2 = \sigma_\omega^2 + \gamma^{-1}\sigma_c^2 q_N + \gamma^{-2}\sigma_c^4\nu^2\tilde{m}^2\sigma_\epsilon^2. \tag{6.82}$$

These are nearly identical to the equations we had before. The two key changes are the presence of a term with a negative sign inside the coefficient σ_N of the random variable z_N, and the $\gamma^{-1}\sigma_c^2\nu\tilde{m}$ term inside the parentheses in the equation for the resources.

We can now proceed in the same way as before, solving for $\bar{N}_0$ and $\bar{R}_0$ and taking derivatives to compute the susceptibilities. We find:

$$\chi = -\frac{1}{2\gamma^{-1}\sigma_c^2\nu}\left\{1 - \left\langle \frac{\omega+\gamma^{-1}\mu\langle N\rangle+\sigma_R z_R+\gamma^{-1}\sigma_c^2\nu\tilde{m}}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle+\sigma_R z_R+\gamma^{-1}\sigma_c^2\nu\tilde{m})^2-4\gamma^{-1}\sigma_c^2\nu(K+\delta K_0)}}\right\rangle\right\} \tag{6.83}$$

$$\nu = -\frac{\phi_N}{\sigma_c^2\chi} \tag{6.84}$$

This is almost the same as the expression in Equation (6.64) obtained in the absence of constraints, except for the extra term $\gamma^{-1}\sigma_c^2\nu\tilde{m}$ in the numerator and denominator. This term is significant because ν is a negative number, and if its absolute value is large enough, it can make the whole term in angle brackets negative. Inserting the second

equation into the first, we obtain a formula for S^*/M:

$$\frac{S^*}{M} = \gamma^{-1}\phi_N \quad = \tfrac{1}{2}\left\{1 - \left\langle \frac{\omega+\gamma^{-1}\mu\langle N\rangle+\sigma_R z_R+\gamma^{-1}\sigma_c^2\nu\tilde{m}}{\sqrt{(\omega+\gamma^{-1}\mu\langle N\rangle+\sigma_R z_R+\gamma^{-1}\sigma_c^2\nu\tilde{m})^2-4\gamma^{-1}\sigma_c^2\nu(K+\delta K_0)}}\right\rangle\right\} \tag{6.85}$$

The term in brackets can now be negative, but is always greater than -1. We thus obtain the bound:

$$\frac{S^*}{M} < 1. \tag{6.86}$$

The term approaches -1 in the limit $\nu \to -\infty$, which is the same limit required to saturate the bound in the model with self-renewing resources. As in that case, the limit cannot actually be achieved, because $\nu \to -\infty$ implies $\chi \to 0$ (Equation (6.84)), and χ appears in the denominator of the final expression for $\bar{N}_0$ (Equation (6.27)), while the numerator always remains finite.

The only way to achieve the limit $\frac{S^*}{M} = 1$ is to make the numerator vanish in the same way as the denominator, which can only happen in the presence of hard constraints $\sigma_m = \sigma_\epsilon = 0$. In this case, it is easy to see that setting $R_\alpha = \tilde{m}$ for all α and $\chi \to 0, \nu \to -\infty$ solves both the steady state equations, regardless of the value of $\tilde{N}_0$. In Equation (6.79) for $\tilde{N}_0$, the mean and the fluctuating part inside the brackets both vanish individually ($\mu\langle R\rangle - \mu\tilde{m} = 0$, $\sigma_N = 0$), and the back-reaction term also vanishes ($\sigma_c^2\chi\bar{N}_0 = 0$), leaving the equation trivially satisfied. In Equation (6.80) only the terms with ν are significant in this limit, and they cancel each other perfectly. This is the "shielded phase" discussed in [TM17].

Note also that if we take the $\chi \to 0$, $\nu \to -\infty$ limit first, before performing any substitutions, Equations (6.83) and (6.84) are satisfied independently of the choice of ϕ_N. This means that $\gamma^{-1}\phi_N = S^*/M$ can be greater than 1, as observed in the simulations of [PTW17].

6.3.4 Numerical evidence

We show a comparison between the cavity solution and numerical results in Fig. 6.6 and Fig. 6.4 for three different distributions of the consumption matrix $\mathbf{C}$. For the Gaussian and Bernoulli distributions, $\frac{S^*}{M}$ can reach the upper bound we derived for two different resource dynamics. For externally supplied resource dynamics, $\frac{S^*}{M}$ never exceeds 0.5. For the uniform case, since the fluctuation of consumption matrix is small, the niche overlap is large and there is fierce competitions among species and theses ecosystems live very far from the upper bounds we derive. However, even for the uniform case, the

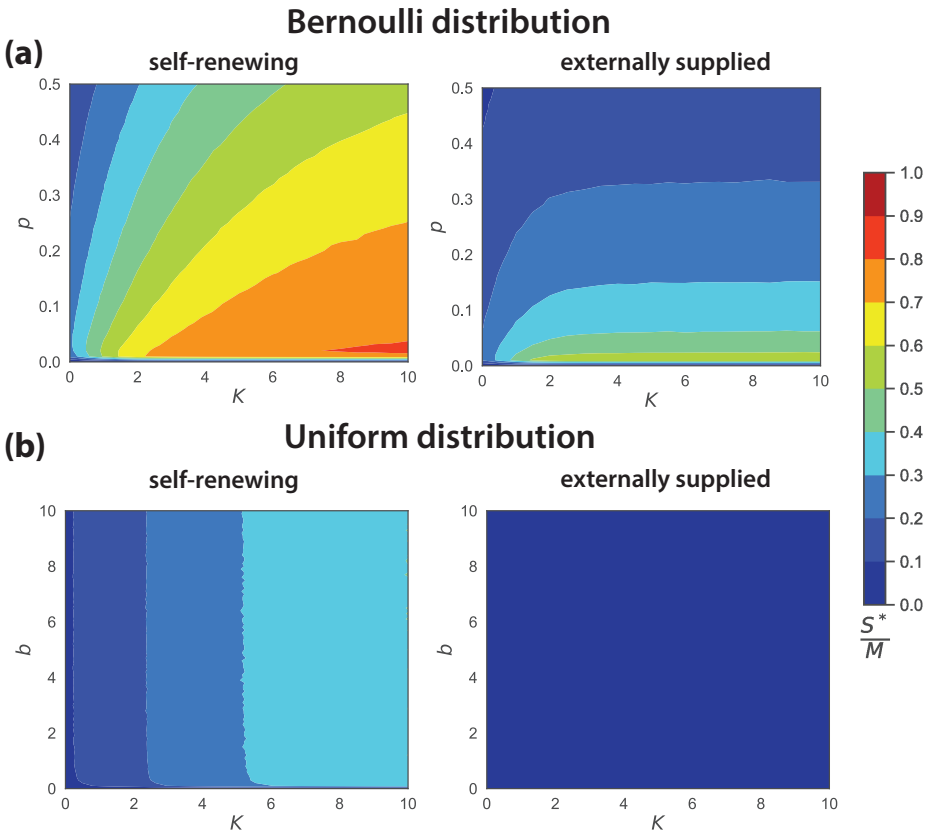

FIGURE 6.4: Comparison of species packing $\frac{S^*}{M}$ for different distributions of consumption matrices $\mathbf{C}$ with self-renewing and externally-supplied resource dynamics. The simulations represent averages from 1000 independent realizations with the system size $M = 100$, $S = 500$ and parameters at the Appendix B.2.

species packing fraction is significantly larger for self-renewing resource dynamics than externally supplied resource dynamics.

6.4 Results

In Chapter 4, we derived a mean-field cavity solution for steady-state dynamics of the the GCRM with self-renewing resource dynamics in the high-dimensional limit where the number of resources and species in the regional species pool is large ($S, M \gg 1$)[ABM18b, MCWMI19, CMIM19]. The overall procedure for deriving the cavity equations for GCRM with externally supplied resource is similar to that for GCRMs with self-renewing resources and is shown in Fig. 6.2. We assume the K_α and m_i are independent random normal variables with means K and m and variances σ_K^2 and σ_m^2, respectively. We also assume ω_α are independent normal variables with mean ω and variance σ_ω^2. The elements of the consumption matrix $C_{i\alpha}$ are drawn independently

from a normal distribution with mean μ/M and variance σ_c^2/M. This scaling with M is necessary to guarantee that $\langle N\rangle$, $\langle R\rangle$ do not vanish when $S, M \gg 1$ with $M/S = \gamma$ fixed. Later, we will consider a slightly modified scenario where the maintenance costs are correlated with the consumption matrix in order to implement the metabolic trade-offs discussed above.

The basic idea behind the cavity method is to derive self-consistency equations relating an ecosystem with M resources and S species to an ecosystem with $M+1$ resources and $S+1$ resources. This is done by adding a new "cavity" species 0 and a new "cavity" resource 0 to the original ecosystem. When $S, M \gg 1$, the effect of the new cavity species/resource is small and can be treated using perturbation theory. The cavity solution further exploits the fact that since the $C_{i\alpha}$ are random variables, when $M \gg 1$ the sum $\sum_\alpha C_{i\alpha}R_\alpha$ will be well described by a by a normal distribution with mean $\mu\langle R\rangle$ and variance $\sigma_c^2 q_R$ where $q_R = \langle R^2\rangle = 1/M\sum_\alpha R_\alpha^2$ (see Appendix for details). Combining this with the non-negativity constraint, the species distribution can be expressed as a truncated normal distribution,

$$\bar{N} = \max\left[0, \frac{\mu\langle R\rangle - m + \sqrt{\sigma_c^2 q_R + \sigma_m^2}z_N}{\sigma_c^2\chi}\right] \tag{6.87}$$

where $\chi = -\left\langle\frac{\partial\bar{R}_\alpha}{\partial\omega_\alpha}\right\rangle = -M^{-1}\sum_\alpha\frac{\partial\bar{R}_\alpha}{\partial\omega_\alpha}$ and z_N is a standard normal variable. This equation describes GCRMs with both externally supplied and self-renewing resource dynamics [ABM18b].

The steady-state cavity equations for externally supplied resources are significantly more complicated and technically difficult to work with than the corresponding equations for self-renewing resources. To see this, notice that the steady-state abundance of resource α can be found by plugging in Eq. 6.3 into Eq 6.1 and setting the left hand side to zero to get

$$\bar{R}_\alpha = K_\alpha/(\omega_\alpha + \sum_j \bar{N}_j C_{j\alpha}) = \frac{K_\alpha}{\omega_\alpha^{\text{eff}}}, \tag{6.88}$$

where we have defined $\omega_\alpha^{\text{eff}} = \omega_\alpha + \sum_j \bar{N}_j C_{j\alpha}$. When $S \gg 1$, both the denominator $\omega_\alpha^{\text{eff}}$ and the numerator K_α can be modeled by independent normal random variables. This implies that the the steady-state resource abundance is described by a ratio of normal variables (i.e. the Normal Ratio Distribution) instead of a truncated Gaussian as in the self-renewing case [M$^+$06]. At large σ_c, this makes solving the cavity equations analytically intractable. Luckily, if the variance of the denominator $\omega_\alpha^{\text{eff}}$ is small compared with the mean – which is true when σ_c not too large – we can still obtain an approximate replica-symmetric solution by expanding in powers of the standard deviation over the mean of $\omega_\alpha^{\text{eff}}$ (see Appendix). We consider expansions to the cavity solutions where the

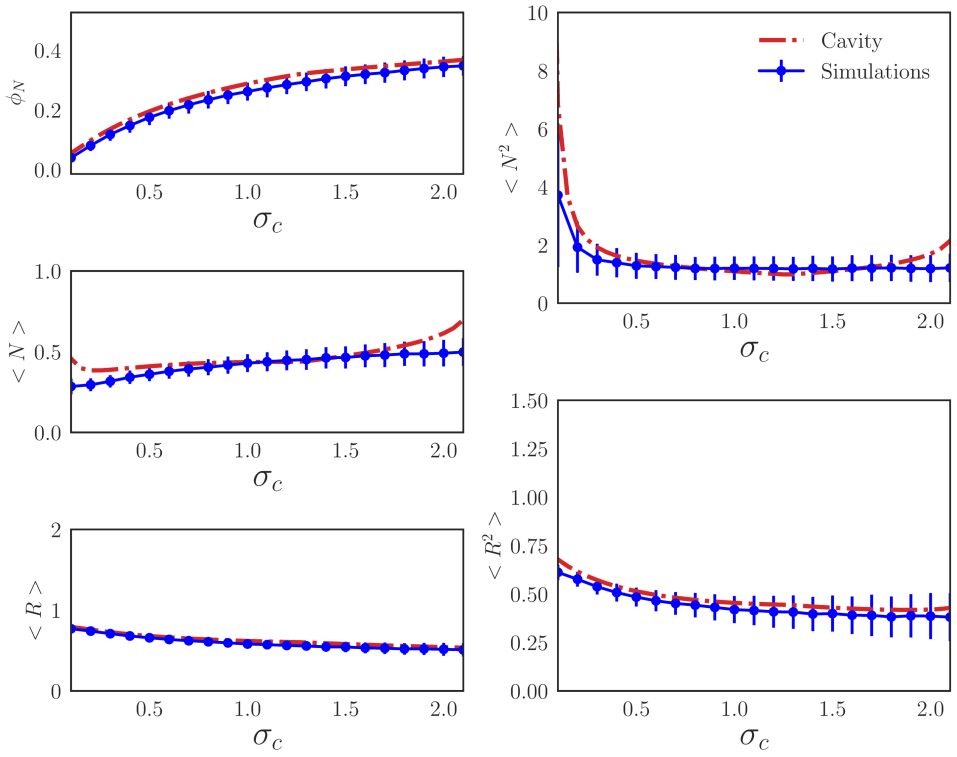

FIGURE 6.5: Comparison between cavity solutions (see main text for definition) and simulations for the fraction of surviving species $\phi_N = \frac{S^*}{S}$ and the first and second moments of the species and resources distributions as a function of σ_c. The error bar shows the standard deviation from 1000 numerical simulations with $M = S = 100$ and all other parameters are defined in the Appendix B.2. Simulations were run using the CVXPY package [AVDB18].

denominator in Eq. 6.88 is expanded to 1^{st} order. In general, the backreaction correction is quite involved since resources and species form loopy interactions resulting in non-trivial correlation between $C_{i\alpha}$ and N_i that must be properly accounted for (see Appendix6.2.5).

6.4.1 Comparison with numerics

The full derivation of 1^{st} order expansions of the mean-field equations are given in the Appendix. The resulting self-consistency equations can be solved numerically in Mathematica. Fig. 6.5 shows a comparison between the cavity solution and 1000 independent numerical simulations for various ecosystem properties such as the fraction of surviving species S^*/S and the first and second moment of the species and resource distributions (simulation details are in the Appendix B.2). As can be seen in the figure, our analytic expressions agree remarkably well over a large range of σ_c. However, at very large σ_c (not shown), the cavity solutions start deviating from the numerical simulations because

the Ratio Normal Distribution can no longer be described using the 1^{st} order expansion to the full cavity equations.

As a further check on our analytic solution, we ran simulations where the $C_{i\alpha}$ were drawn from different distributions. One pathology of choosing $C_{i\alpha}$ from a Gaussian distribution is that when σ_c is large, many of consumption coefficients are negative. To test whether our cavity solution still describes ecosystems when $C_{i\alpha}$ are strictly positive, we compare our cavity solution to simulations where the $C_{i\alpha}$ are drawn from a Bernoulli or uniform distribution. As before, there is remarkable agreement between analytics and numerics (see Fig. 6.3)

6.4.2 Species packing without metabolic tradeoffs

The essential ingredients needed to derive species packing bounds for GCRMS are the cavity equations for the average local susceptibilities $\nu = \left\langle \frac{\partial \bar{N}_i}{\partial m_i} \right\rangle = S^{-1} \sum_j \frac{\partial \bar{N}_i}{\partial m_i}$ and $\chi = \left\langle \frac{\partial \bar{R}_\alpha}{\partial X_\alpha} \right\rangle = M^{-1} \frac{\partial \bar{R}_\alpha}{\partial X_\alpha}$, with $X_\alpha = K_\alpha$ for externally supplied resources and $X_\alpha = -\omega_\alpha$ for self-renewing resources. These two susceptibilities measure how the mean species abundance and mean resource abundance respond to changes in the species death rate and the resource supply/depletion rate, respectively. They play an essential role in the cavity equation and can be used for distinguishing different phases in complex systems[RMS15, CMIM19].

For the self-renewing case, the susceptibilities χ_s and ν_s are given by eq. (59, 60) in [MCWMI19]

$$\nu_s = -\frac{\phi_N}{\sigma_c^2 \chi_s}, \quad \chi_s = \frac{\phi_R}{1 - \gamma^{-1}\sigma_c^2 \nu_s}, \tag{6.89}$$

and can be reduced to $\chi_s = \phi_R - \gamma^{-1}\phi_N$, where $\phi_R = M^*/M$, with M^* equal to the number of non-extinct resources in the ecosystem. In order to guarantee the positivity of $\langle N \rangle$, we must have $\chi_s = \phi_R - \gamma^{-1}\phi_N > 0$, resulting in an upper bound

$$1 \geq \frac{M^*}{M} > \frac{S^*}{M} \tag{6.90}$$

which states that the number of surviving resources must be smaller than the number of surviving species.

For the externally supplied case, the corresponding equations take the form

$$\nu = -\frac{\phi_N}{\sigma_c^2 \chi}, \chi = -\frac{1}{2\gamma^{-1}\nu\sigma_c^2}(1 - \langle ... \rangle), \tag{6.91}$$

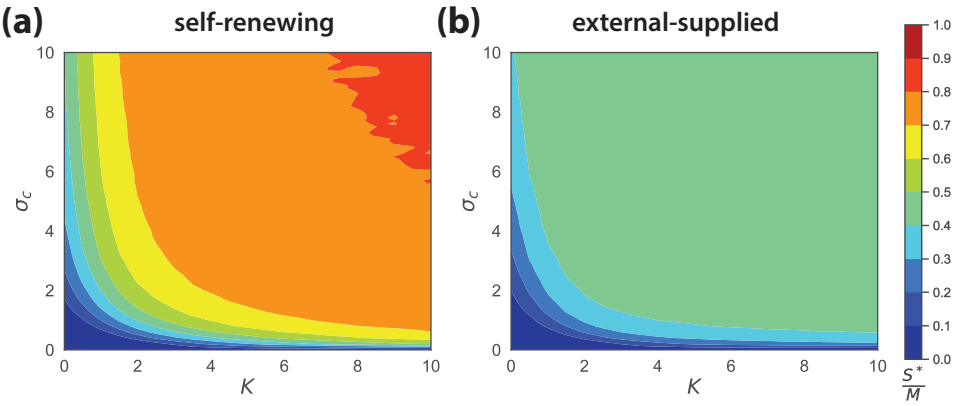

FIGURE 6.6: Comparison of the species packing ratio $\frac{S^*}{M}$ at various σ_c and K for self-renewing and externally supplied resource dynamics. The simulations represent averages from 1000 independent realizations with the system size $M = 100$, $S = 500$ (parameters in Appendix B.2).

where the full expression of $\langle ... \rangle$ can be found in eq. (6.64) in the appendix. For our purposes, the most important property is that in the absence of metabolic tradeoffs, the expression $\langle ... \rangle$ is always *positive*. Combining this observation with the equations above gives the upper bound

$$\frac{1}{2} > \frac{S^*}{M} = \phi_N \gamma^{-1}. \tag{6.92}$$

Thus, for externally supplied resources, at most *half of all potential niches* are occupied. Fig. 6.6 shows numerical simulations confirming the species packing bound for various choices of K and σ_c.

6.4.3 Species packing with metabolic tradeoffs

We also find that metabolic tradeoffs modify the cavity equations in such a way that the expression in brackets $\langle ... \rangle$ in Equation (6.91) can become negative (see Appendix). However, it still remains greater than -1, allowing us to derive a species packing bound of the form $S^* < M$ even in the presence of soft metabolic constraints. In Figure 6.7, we simulated various ecosystems where the maintenance costs of species were chosen to obey metabolic tradeoffs of the form $m_i = \sum_\alpha C_{i\alpha} + \delta m_i$, where δm_i are i.i.d. normal variables with variance σ_m^2. Note that a larger σ_m corresponds to ecosystems with softer metabolic constraints. We found that when $\sigma_m/\sigma_c > 1$, these ecosystems obey the $1/2$ species packing bound derived above. This can also be analytically shown using the modified cavity equations derived in the appendix. Finally, we show in the appendix that when the metabolic tradeoffs take the form of hard constraints on the consumer preferences as in [AF19, TM17, PTW17, LLL+19], the cavity equations allow for interesting non-generic behavior with $S^* \geq M$, consistent with these previous works. Importantly, we

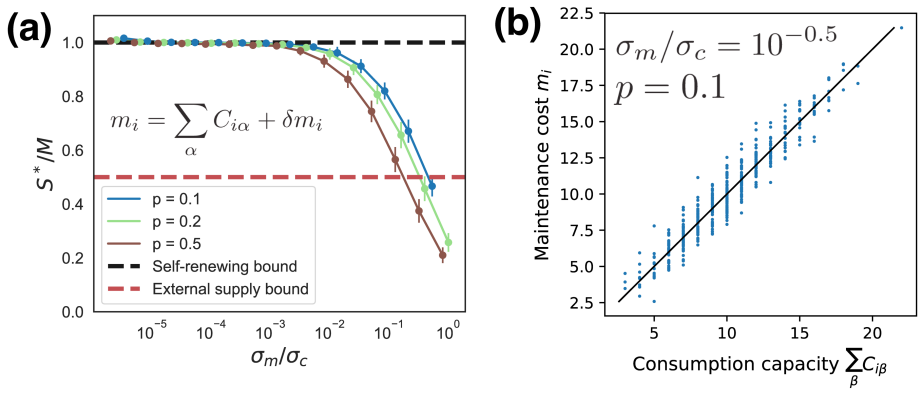

FIGURE 6.7: Species packing bounds in the presence of metabolic tradeoffs. (a) The species packing ratio S^*/M as a function of σ_m/σ_c, where σ_m is the standard deviation of the δm_i and $\sigma_c/\sqrt{M}$ is the standard deviation of $C_{i\alpha}$. Simulations are for binary consumer preference matrix $C_{i\alpha}$ drawn from a Bernoulli distribution with probability p. (b) m_i versus $\sum_\alpha C_{i\alpha}$ for $p = 0.1$ and $\sigma_m/\sigma_c = 10^{-0.5}$ See Appendix for all parameters.

find that even modest modifications of the tradeoff equation $m_i \propto \sum_\alpha C_{i\alpha}$ results in ecosystems that satisfy the $1/2$ species packing bound.

6.4.4 Classifying ecosystems using species packing

Recently, it has become clear that there is a deep relationship between ecosystem and constraint satisfaction problems [MCWMI19, MICM20, TM17, AF19]. In particular, each species can be thought of as a constraint on possible resource abundances [MCWMI19, MICM20]. Inspired by jamming [LN10] , this suggests that we can separate ecosystems into qualitatively distinct classes depending on whether the competitive exclusion bound is saturated. We designate ecosystems where $S^* \to M$ (like GCRMs with self-renewing resources) as *isostatic species packings*, and ecosystems where the upper bound $S_{\max}$ on the number of surviving species is strictly less than the number of resources $S^* < S_{\max} < M$ (like GCRMs with externally supplied resources without metabolic tradeoffs) as *hypostatic species packings*. Ecosystems with $S^* \geq M$ (like GCRMs with hard metabolic constraints) are designated as *non-generic species packings* because of the presence of a macroscopic number of additional hards constraints (i.e. the number of additional constraints that are imposed scales with S and M in the limit $S, M \to \infty$). This basic schema suggests a way of refining the competitive exclusion principle and may help shed light on controversies surrounding the validity of basic species packing bounds.

6.5 Discussion

In this Chapter, we examine the effect of resource dynamics on community structure and large-scale ecosystem level properties. To do so, we analyzed generalized Consumer Resource Models (GCRMs) with two different resource dynamics: externally supplied resources that are supplied and degraded at a constant rate and self-replicating resources whose behavior in the absence of consumers is well described by a logistic growth law. Using a new cavity solution for GCRMs with externally supplied resources and a previously found cavity solution of the GCRM with self-renewing resources, we show that the community structure is fundamentally altered by the choice of resource dynamics. In particular, for externally supplied resources, we find that species generically can only occupy *half* of all available niches whereas for self-renewing resources all environmental niches can be filled. We confirm this surprising bound using numerical simulations.

These results show how resource dynamics, which are neglected in commonly used Lotka-Volterra models, can fundamentally alter the properties of ecosystems. Much work still needs to be done to see if and how our results must be modified to account for other ecological processes such as demographic stochasticity, spatial structure, and microbe-specific interactions such as cross-feeding [GLB+18, MICG+19]. It will also be necessary to move beyond steady-states and consider the dynamical properties of these ecosystems. More generally, it will be interesting to further explore the idea that we can classify ecosystems based on species-packing properties and see if such a schema can help us better understand the origins of the incredible diversity we observe in real-world ecosystems.

Chapter 7

Summary and future directions

Nature has revealed an astounding degree of phylogenetic and physiological diversity in microbial communities. The recent advances in DNA sequencing technologies have resulted in the generation of large amounts of microbial data. The ability to measure microbial community species abundances with high resolution has opened a precision era in microbial ecology. Understanding such a large amount of microbial data challenges current theories and analytical approaches. Historically, theoretical ecologists have devoted considerable effort to analyze ecosystems consisting of a few species. However, analytical approaches and theoretical insights derived from small ecosystems may not scale up to large ecosystems. On the other hand, most ecological data are high dimensional, involving hundreds of species across different habitats. A huge number of combinations of states must be considered, and the use of exhaustive search strategies to exactly determine interactions among species is no longer feasible, i.e., *the curse of dimensionality* [Ric57].

As we all know, statistical mechanics has allowed us to quantitatively describe collective phenomena in solids and plasma, large-scale structures in astrophysics, and understand how collective behaviors emerge from the interaction of many individual components[Ma18]. In the past few years, I have built a theoretical framework with statistical mechanics, inspired by spin-glass theory, to understand community structures and coexistence patterns in complex ecosystems. I have also developed computational tools to model microbial ecosystems and analyze high-dimensional microbial data.

7.1 Spin Glasses and Ecology

My Ph.D. thesis work mainly focused on using the cavity method, to answer fundamental questions in ecology. The basic idea behind the cavity method is to derive

self-consistency equations by relating an ecosystem with S species / M resources to another ecosystem with S+1 species / M+1 resources at the thermodynamic limit. When $S, M \to \infty$, there is no difference between statistical observables computed in both disorder systems. Adding a new "cavity" species/resource 0 to the original ecosystem results in a perturbation to the original equilibrium state. We can derive a couple of self-consistency equations by relating the original state to the perturbed state.

In [CMIM19], we explored the effects of noise on diverse communities with designed structures. This problem can be traced to Robert May's pioneer work in 1972 about the stability of large complex ecosystems [May72]. Using the circular law of large random matrices derived by Giniber [Gin65], May showed that species can have unbounded growth when the amplitude of the noise is compatible with the intra-specific interactions. In May's model, all ecosystem properties are encoded in the species-species interaction matrix. A major limitation of these models is that they neglect resources, making it difficult to understand how ecosystem properties depend on both the external environment and species consumer preferences.

Our work focused on MacArthur's Consumer Resource Model (CRM) where species are modeled as consumers that can consume resources [ML67a]. The bipartite nature of the CRM, from the presence of two types of degrees of freedom: resources and species, results in a Wishart matrix [RCKT08]. Random Wishart matrices are well-known to follow the Marchenko-Pastur law [MP67b]. The designed structure with noise in ecology can be classified as a class of "signal + noise" problems, which can be described in the framework of spiked random matrix models [BS06]. We showed that there is a threshold value for the strength of specific interactions over the amplitude of noise s, similar to the signal-to-noise ratio, below which, ecological properties of communities are indistinguishable from purely disorder ecosystems.

In addition to its theoretical significance, our results also yield interesting biological implications. First, it suggests that bottom-up engineering of complex ecosystems may be very difficult. As the number of components increases, small uncertainties in each of the interaction parameters may eventually overwhelm the designed interactions, and destabilize the intended steady state. Instead, such system are much more likely end up in a typical state, which our theory suggests is much more stable than the intended designed state as ecosystems become more diverse. Second, our results may also help explain the surprising success of consumer-resource models with random parameters in predicting the behavior of microbial ecosystems in the lab and natural environments [GLB+18, MCM20]. They also suggest the possibility of predicting macroscopic ecosystem-level properties like diversity or total biomass even when ecosystems are poorly characterized or have lots of missing data.

In [CMIM20], we developed cavity method to calculate the capacity of random ecosystems under various biological constraints. The competitive exclusion principle asserts that coexisting species must occupy distinct ecological niches (i.e., the number of surviving species can not exceed the number of resources) [MA77]. While recent study shows strict metabolic trade-off can result in the coexistence of infinity number of species with only few numbers of resources provided [TPMW17]. Our work resolves these controversies by showing that generically, the competitive exclusion principle holds and that one needs very fine-tuned and strict metabolic tradeoffs to violate this principle. Another major contribution of the work is to show that dynamics of resources in an ecosystem can qualitatively change the number of species that can survive in an ecosystem: for resources that are self-renewing (can reproduce) the number of species that survive in an ecosystem can approach the number of resources; in contrast, if resources that are externally supplied, the number of species that can survive is bounded by one-half the number of resources in the environment. Thus, in this latter case, the competitive exclusion bound is never saturated. Besides the ecological significance of our results, the problem we solved is equivalent to a machine learning problem: how many random linear constraints are active in optimizing quadratic and Kullback-Leibler(KL) divergence objective functions. Our result shows the number of active constraints in the KL-divergence case is only equal to half of the number in the quadratic case.

Cavity method we used is still in the replica-symmetry (unique attractor) regime. [BBC18b] and [Bun17] show, in simple Lotka-Volterra models, the replica-symmetry-breaking phenomenon emerges and there is the phase transition from unique to multiple attractor phase. In the multiple-attractor phase, the dynamics system is marginally stable and highly sensitive to initial conditions, similar to the de Almeida-Thouless line in spin glass theories [dAT78] and chaotic behavior in random one-layer neural networks [SCS88]. However, our replica symmetric cavity method fails in the multiple-attractor phase and it becomes tedious to consider multiple-attractor corrections. In the future, I am very interested in expanding our theoretical framework to the multiple equilibria regime. I feel this can serve as a scaffold to understand how living organisms perform complex behaviors, such as assemble patterns in microbial communities [MICG+19], or even understand mixed selectivity in neuron science [RBW+13].

In summary, we have developed sophisticated theories about complex ecosystems. I believe our methodology is not limited to ecology. I want to apply and generalize these statistical-physics approaches to study other biological systems, including quality control in protein synthesis, single-cell transcriptional dynamics, gene regulation in cell divisions, and effects of drug combinations.

7.2 Inferences in Ecology

In additional to purely theoretical projects, I am also interested in developing inference tools to analyze high-dimensional microbial data. There are natural connections between statistical mechanics and inference. The cavity method is equivalent to belief propagation (BP) algorithms used in optimization and machine learning. In particular, the cavity equations can be naturally thought of as a message-passing algorithm for performing inference on graphical models [YFW01]. BP has been widely applied to solving inference problems in many scientific fields, including protein structure[CFF+18], Boltzmann machines [WT03], random K-SAT [MPZ02], compressed sensing [KMS+12] and other combinatorial optimization problems [ZK16]. The last section shows after specifying the interaction components, the cavity method is successful in predicting coexistence patterns, especially abundance distributions, and identifying assembly rules for microbial communities. However, we can also invert this logic, and ask if we can use these methods to infer interactions from microbial population abundance data.

Aside from the intrinsically high dimensional nature of microbial abundance data, this inference task is challenging for several additional reasons. First, empirical biological networks show that the interacting components in ecosystems are often are sparse and structured [All20], making it difficult to use mean-field approximations. Second, biological measurement is always accompanied by noise, especially when using high throughput single-cell technologies [ABBR18]. It is essential to filter noise before inferring meaningful signals. Third, sequencing data are inherently relative. The relative abundances can be perfectly negatively correlated, even though the correlation of their absolute values is not related [PGB11].

Several generative model-based frameworks have been proposed to address above issues partially [FMW17, YSL+19]. After specifying the mathematical model of dynamical biological processes, model parameters are determined by minimizing the loss function between model predictions and experimental data with gradient descent. As a result, the choice of mathematical model is a critical component of any inference procedure. However, most mathematical models used for inference only consider pairwise interactions. As ecosystems gets large, the number of parameters grow rapidly, especially if high-order terms are needed. As a result, the inference model becomes too complex and it is difficult to avoid overfitting. To circumnavigate these problems, I would like to use functional structures of microbial communities to reduce model complexity while still allowing for computationally feasible inference methods. For example, we developed classic MacArthur consumer-resource model including crossfeeding interactions, whose metabolic interaction components result from biochemical pathways and can be estimated from flux balance analysis [MICG+19].

I also would like to improve mean-field inference approaches for understanding deep neural networks (DNN). Mean-field approximation assumes a unimodal prior distribution, and thus never works for multi-modal posterior distributions accurately, a scenario which is likely to be common in high-dimensional biological systems. DNNs are well-suited to learning functional approximation of high-dimensional data. Thus, we can use DNN as a more expressive ansatz replacing the Bethe approximation in mean-field theories (MFT). For example, graph neural network (GNN) employ the message-passing algorithm and DNN to estimate the node representation from its neighbors. Because the setup of MFT varies in different models, the structure of the corresponding GNNs should also be changed when solving different problems. One promising direction is to combine GNNs with statistical-physics-inspired approaches, such as belief propagation, survey propagation, to solve random K-SAT and other related problems. I believe this direction can lead to novel inference algorithms for biological data analysis.

Appendix A

Basic material on Lotka-Volterra model

Lotka-Volterra equations are foundational phenomenological; here we collect basic mathematical facts useful for analyzing such equations for stability and dynamics. Lotka-Volterra dynamics with S species is described by equation A.1.

$$\frac{dN_i}{dt} = N_i(r_i - \sum_{i \neq j} A_{ij} N_j), \quad i = 1, 2, \ldots S \tag{A.1}$$

A.1 A proxy for the Jacobian

A common question is to understand the local stability of fixed points of this equation, i.e., if the ecosystem will return to the same fixed point after a small perturbation of population abundance N_i.

A direct approach to local is to compute the fixed point of the above equations $\bar{N} = A^{-1}r$, expand to linear order about that fixed point $N = \bar{N} + \delta N$ and thus determine the Jacobian about that fixed point, $J_{ij} = \bar{N}_i A_{ij}$. The fixed point is stable if all the eigenvalues of J_{ij} are negative.

However, such a criterion is not convenient since it involves computing the fixed point abundances $\bar{N}$. Thus some effort has been expended in finding a simpler stability criterion in terms of the competition matrix A_{ij}, instead of examining the actual Jacobian $J_{ij} = \bar{N}_i A_{ij}$.

One such simple sufficient condition for stability is that $A + A^T$ be negative definite. This simple rule can be derived using results on D-stability[KB12]: A real square matrix

$\mathbf{A}$ is said to be D-stable if the matrix $\mathbf{DA}$ is negative definite for every choice of a positive diagonal matrix $\mathbf{D}$. A sufficient condition for D-stability is that $\mathbf{A} + \mathbf{A^T}$ is negative definite(using lemma 2.1.4 in [KB12]). Thus, if the Lotka-Volterra competition matrix $\mathbf{A}$ is such that $A + A^T$ is negative definite, we can conclude immediately that A is D-stable and hence $J = \bar{N}A$ is negative definite, assuming all steady state abundances $\bar{N}$ are non-negative.

We present an alternative proof of this sufficient condition, i.e., if $\mathbf{A} + \mathbf{A}^T$ is negative definite, for any $\bar{\mathbf{N}}$ with all $N_i > 0$, $\bar{N}_i A_{ij}$ is also negative definite.

$$\mathbf{M} = \bar{\mathbf{N}}^{1/2}(\bar{\mathbf{N}}^{1/2}\mathbf{A}\bar{\mathbf{N}}^{1/2}))\bar{\mathbf{N}}^{-1/2} \tag{A.2}$$

where $\bar{\mathbf{N}}^{1/2}$ is the diagonal matrix whose entries are the square roots of the population sizes. This equation says that $\mathbf{M}$ is similar to $\mathbf{W} \equiv \bar{\mathbf{N}}^{1/2}\mathbf{A}\bar{\mathbf{N}}^{1/2}$, which implies that they share the same eigenvalues. Since $\mathbf{W}$ and $\mathbf{A}$ are both symmetric matrices, their eigenvalues are all real, and the positivity of all the eigenvalues is equivalent to the negative-definiteness of the matrix.

Now we note that $\mathbf{W}$ is negative definite if and only if $\mathbf{A}$ is negative definite. For if $\mathbf{A}$ is negative definite, then $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$ for all column vectors $\mathbf{x} \neq 0$, including the column vector $\mathbf{x} = \bar{\mathbf{N}}^{1/2}\mathbf{y}$ for any column vector $\mathbf{y} \neq 0$. But this implies that $\mathbf{y}^T\bar{\mathbf{N}}^{1/2}\mathbf{A}\bar{\mathbf{N}}^{1/2}\mathbf{y} < 0$ for all $\mathbf{y} \neq 0$, i.e., that $\mathbf{W}$ is positive definite. Conversely, if $\mathbf{W}$ is positive definite, then $\mathbf{y}^T\bar{\mathbf{N}}^{1/2}\mathbf{A}\bar{\mathbf{N}}^{1/2}\mathbf{y} > 0$ for all $\mathbf{y} \neq 0$, including $\mathbf{y} = \bar{\mathbf{N}}^{-1/2}\mathbf{x}$ for any $\mathbf{x} \neq 0$. But this implies that $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \neq 0$, i.e., that $\mathbf{A}$ is negative definite. We conclude that the eigenvalues of $\mathbf{W}$ are all negative if and only if the eigenvalues of $\mathbf{A}$ are all negative.

While this stability result in terms of A_{ij} alone is convenient, note that the stability of $A + A^T$ is not a necessary condition for stability of the Lotka-Volterra equation. That is, $J = \bar{N}A$ can be stable even if $A + A^T$ has positive eigenvalues. As a simple example, consider $A = \left(\begin{smallmatrix} 2 & -2/c \\ c & -1 \end{smallmatrix}\right)$ for sufficiently large c. The eigenvalues of A and $A + A^T$ are all positive, indicating instability; indeed, the positive self-interaction of species 1 also suggests instability. However, the steady state abundances of this matrix are .. . Hence A generates stable LV dynamics at one of the fixed points, even though $A + A^T$ has negative eigenvalues.

Further discussions about D-stability and its application in ecology can be found in [SCG+18].

A.2 Structural stability

A related question to stability is the robustness of a stable fixed point to perturbations in parameters such r_i, A_{ij}, e.g., due to environmental perturbations. Structural stability is one such measure of such robustness to changes in the carrying capacity values r_i. We first define a fixed point $\bar{\mathbf{N}}$ to be 'feasible' if all S species in the ecosystem can coexist, i.e., $\bar{N}_i = \sum_j A_{ij}^{-1} r_j > 0$.

Structural stability is then defined as the fractional volume Ω of all (normalized) vectors r_i such that $\bar{N}_i = \sum_j A_{ij}^{-1} r_j > 0$:

$$\Omega = \frac{2^S \int_{-\infty}^{\infty} \Pi_{i=1}^{S} dr_i \delta(||\mathbf{r}|| - 1) \Pi_{h=1}^{S} \Theta(\sum_j A_{hj}^{-1} r_j)}{\int_{-\infty}^{\infty} \Pi_i dr_i \delta(||\mathbf{r}|| - 1)}. \tag{A.3}$$

The factor 2^S is for normalization as $\Omega = 1$ when all off-diagonal elements of $\mathbf{A}$ are zero, i.e., there is no interaction between species. The structural stability Ω has a natural geometric explanation that it is a high-dimensional solid angle for a convex cone defined by $\left\{\mathbf{r} \in \mathbf{R}^S | \sum_j A_{hj}^{-1} r_j > 0\right\}$ [GS10].

Precise estimation of the integral eq. (A.3) can only be done for a ecosystem consisting of few species. For complex random ecosystems, it can be estimated approximately with mean field theories. Technical details can be found in [GAS+17]. For consumer-resource dynamics, the mathematical expression for Θ is different and can be found in [BO18].

Note that structural stability is defined in terms of positive steady state $\bar{N}_i$ In local stability analysis, a feasible fixed point must be assumed before adding the perturbations. In the structural stability calculation, $\mathbf{A}$ is invertible and for an ecosystem lying exactly on May's stability criteria, $\mathbf{A}$ is not invertible. We can see there is a close connection between stability and feasibility and it need to carefully address their relations in theoretical analysis. Further discussion can be found in [RSB14, DVR+18].

Appendix B

Simulation details

B.1 Chapter 5

All simulations are done with the CVXPY package[AVDB18] in PYTHON 3. All codes are available on GitHub at https://github.com/Emergent-Behaviors-in-Biology/typical-random-ecosystems.

B.1.1 Parameters

- Figure 5.1(B), 5.2(B), 5.3(C, D): the consumer matrix $\mathbf{C}$ is sampled from the Gaussian distribution $\mathcal{N}(\frac{\mu}{M}, \frac{\sigma_c}{\sqrt{M}})$. $S = 100$, $M = 100$, $\mu = 0$, $K = 1$, $\sigma_K = 0.1$, $m = 0.1$, $\sigma_m = 0.01$, and each data point is averaged from 5000 independent realizations. The model is simulated with eqs. (5.2).

- Figure 5.2(C): the consumer matrix $\mathbf{C}$ is sampled from the uniform distribution $\mathcal{U}(0, b)$. $S = 100$, $M = 100$, $\mu = 0$, $K = 1$, $\sigma_K = 0.1$, $m = 0.1$, $\sigma_m = 0.01$, and each data point is averaged from 5000 independent realizations. The model is described by eqs. (5.2).

- Figure 5.2(D): the consumer matrix $\mathbf{C}$ is sampled from the Bernoulli distribution $Bernoulli(p_c)$. $S = 100$, $M = 100$, $\mu = 0$, $K = 1$, $\sigma_K = 0.1$, $m = 0.1$, $\sigma_m = 0.01$, and each data point is averaged from 5000 independent realizations. The model is described by eqs. (5.2).

- Figure 5.3(B): the simulation is the same as Fig. 5.2(B). Each spectrum is drawn from 10000 independent realizations.

- Figure 5.4: the consumer matrix $\mathbf{C}$ is sampled from the Gaussian distribution $\mathcal{N}(\frac{\mu}{M}, \frac{\sigma_c}{\sqrt{M}})$. $S = 100$, $M = 100$, $\mu = 0$, $K = 1$, $\sigma_K = 0.1$, $m = 0.1$, $\sigma_m = 0.01$.

The model without resource depletion simulated with eqs. (5.2), and each data point is averaged from 5000 independent realizations.. The model with resource depletion is simulated with eqs. (5.1), and each data point is averaged from 4000 independent realizations. Each spectrum is drawn from 1 independent realizations for $S = 500$.

- Figure B.1 : the simulation is the same as Fig. 5.2(B). Each histogram is drawn from 10000 independent realizations.

B.1.2 Distinction between extinct and surviving species

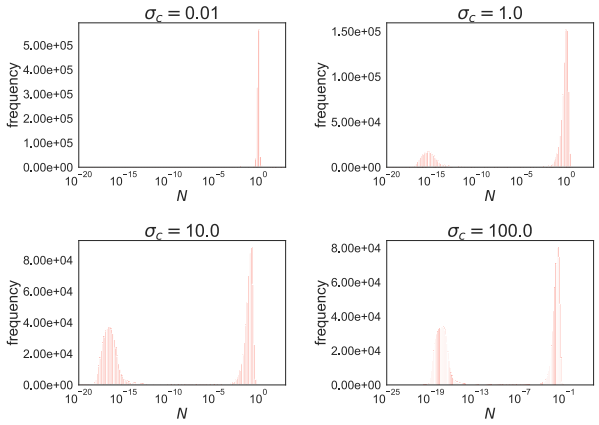

FIGURE B.1: Species abundance N in equilibrium at different σ_c. The simulation details can be found at Appendix B.1.

In the main text, we show that the value of species packing $\frac{S^*}{M}$ in Fig. 5.1 and Fig. 5.2. However, in numerical simulations, even for the extinct species, the abundance is never exactly equal 0 due to numerical errors. As a result, we must choose a threshold to distinguish extinct and surviving species in order to calculate S^*. Since we are using the equivalence with convex optimization to solve the generalized consumer-resource models[MICM20, MCWMI19], we can easily choose a reasonable threshold (e.g. 10^{-10} for both species since the surviving species are well separated in two peaks (see Fig. B.1).

B.2 Chapter 6

All simulations are done with the CVXPY package[AVDB18] in PYTHON 3. All codes are available on GitHub at https://github.com/Emergent-Behaviors-in-Biology/species-packing-bound.

B.2.1 Parameters

- Fig. 6.5: the consumer matrix $\mathbf{C}$ is sampled from the Gaussian distribution $\mathcal{N}(\frac{\mu}{M}, \frac{\sigma_c}{\sqrt{M}})$. $S = 100$, $M = 100$, $\mu = 1$, $K = 1$, $\sigma_K = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$ and each data point is averaged from 1000 independent realizations. We only provide the cavity solution with backreaction here.

- Fig. 6.6 and Fig. B.2: the consumer matrix $\mathbf{C}$ is sampled from the Gaussian distribution $\mathcal{N}(\frac{\mu}{M}, \frac{\sigma_c}{\sqrt{M}})$. $S = 500$, $M = 100$, $\mu = 1$, $\sigma_\kappa = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$ for externally supplied resource dynamics and $S = 500$, $M = 100$, $\mu = 1$, $\sigma_\kappa = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\tau = 1$, $\sigma_\tau = 0$ for the self-renewing one. Each data point is averaged from 1000 independent realizations. For Fig. B.2 , $K = 10$.

- Fig. 6.7: the consumer matrix $\mathbf{C}$ is sampled from the Bernoulli distribution $Bernoulli(p)$ and p are fixed to 0.1, 0.2 and 0.1. m_i follows metabolic tradeoffs Eq. (6.72) with $\sigma_\epsilon = 0$, $\tilde{m} = 1$. We also set $S = 500$, $M = 100$, $K = 10$, $\sigma_K = 0.1$. Each data point is averaged from 100 independent realizations.

- Fig. 6.3(a): the simulation is the same as Fig. 6.5. We show both the cavity solutions with and without reaction here.

- Fig. 6.3(b): the consumer matrix $\mathbf{C}$ is sampled from the Bernoulli distribution $Bernoulli(p)$. $S = 100$, $M = 100$, $K = 1$, $\sigma_K = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$ and each data point is averaged from 1000 independent realizations. The cavity solution is obtained by approximating the Bernoulli distribution to the corresponding Gaussian distribution *i.e.* $\mu = pM$, $\sigma_c = \sqrt{Mp(1-p)}$

- Fig. 6.3(c): the consumer matrix $\mathbf{C}$ is sampled from the uniform distribution $\mathcal{U}(0, b)$. $S = 100$, $M = 100$, $K = 1$, $\sigma_K = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$ and each data point is averaged from 1000 independent realizations. The cavity solution is obtained by approximating the uniform distribution to the corresponding Gaussian distribution, *i.e.* $\mu = bM/2$, $\sigma_c = b\sqrt{M/12}$.

- Fig. 6.4(a): the consumer matrix $\mathbf{C}$ is sampled from the Bernoulli distribution $Bernoulli(p)$. $S = 500$, $M = 100$, $K = 1$, $\sigma_K = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$ and each data point is averaged from 1000 independent realizations. The cavity solution is obtained by approximating the Bernoulli distribution to the corresponding Gaussian distribution *i.e.* $\mu = pM$, $\sigma_c = \sqrt{Mp(1-p)}$

- Fig. 6.4(b): the consumer matrix $\mathbf{C}$ is sampled from the uniform distribution $\mathcal{U}(0, b)$. $S = 500$, $M = 100$, $K = 1$, $\sigma_K = 0.1$, $m = 1$, $\sigma_m = 0.1$, $\omega = 1$, $\sigma_\omega = 0$

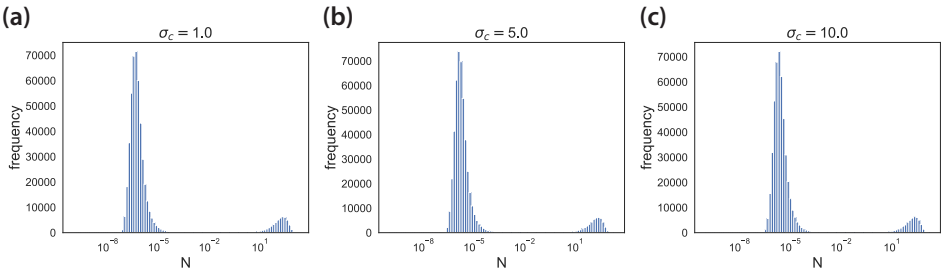

FIGURE B.2: Species abundance N in equilibrium at different σ_c for externally supplied resource dynamics at $K = 10$. The simulations parameters can be found at the Appendix: B.2.

and each data point is averaged from 1000 independent realizations. The cavity solution is obtained by approximating the uniform distribution to the corresponding Gaussian distribution, *i.e.* $\mu = bM/2$, $\sigma_c = b\sqrt{M/12}$.

B.2.2 Distinction between extinct and surviving species

In the main text, we show that the value of species packing $\frac{S^*}{M}$ for the externally supplied resources must be smaller than 0.5. However, in numerical simulations, even for the extinct species the abundance is never exactly equal 0 due to numerical errors. As a result, we must choose a threshold to distinguish extinct and surviving species in order to calculate S^*. Since we are using the equivalence with convex optimization to solve the generalized consumer-resource models[MCWMI19, MICM20], we can easily choose a reasonable threshold (e.g. 10^{-2} in Fig. B.2) since the extinct and surviving species are well separated in two peaks (see Fig. B.2).

Appendix C

Publications List

Identifying feasible operating regimes for early T-cell recognition: The speed, energy, accuracy trade-off in kinetic proofreading and adaptive sorting, **W Cui**, Pankaj Mehta. PloS one 13.8 (2018): e0202331.

Available energy fluxes drive a phase transition in the diversity, stability, and functional structure of microbial communities, Robert Marsland III, **W Cui**, Joshua Goldford, Alvaro Sanchez, Kirill Korolev, Pankaj Mehta. PLoS Comput Biol 15.2 (2019): e1006793 (Chapter 1).

Constrained optimization as ecological dynamics with applications to random quadratic programming in high dimensions, Pankaj Mehta, **W Cui**, Ching-Hao Wang, Robert Marsland III. Phys. Rev. E 99.5 (2019): 052111 (Chapter 2).

A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns, Robert Marsland III, **W Cui**, Pankaj Mehta. Sci Rep 10, 3308 (2020) (Chapter 2).

The Community Simulator: A Python package for microbial ecology, Robert Marsland III, **W Cui**, Joshua Goldford, and Pankaj Mehta. Plos one 15.3 (2020): e0230430 (Chapter 2).

The Minimum Environmental Perturbation Principle: A New Perspective on Niche Theory, R Marsland III, **W Cui**, Pankaj Mehta. The American Naturalist 196.3 (2020): 291-305 (Chapter 2).

Machine Learning as Ecology, Owen Howell, **W Cui**, Robert Marsland III, and Pankaj Mehta. J. Phys. A: Math. Theor. 53 (2020): 334001.

When will complex ecosystems behave like random systems? **W Cui**, Robert Marsland III, Pankaj Mehta. arXiv:1904.02610 (Chapter 5).

Effect of Resource Dynamics on Species Packing in Diverse Ecosystems, Phys. Rev. Lett. 125.4 (2020): 048101. **W Cui**, Robert Marsland III, Pankaj Mehta. Editor's Suggestion, See also the synopsis in Physics Magazine: *Resource Dynamics Dictate Diversity*. (Chapter 6)

The Perturbative Resolvent Method: spectral densities of random matrix ensembles via perturbation theory, **W Cui**, Jason W. Rocks, and Pankaj Mehta. arXiv:2012.00663.

Bibliography

[AABF19] Elena Agliari, Francesco Alemanno, Adriano Barra, and Alberto Fachechi. On the marchenko–pastur law in analog bipartite spin-glasses. *Journal of Physics A: Mathematical and Theoretical*, 52(25):254002, 2019.

[ABBR18] Luis Aparicio, Mykola Bordyuh, Andrew J Blumberg, and Raul Rabadan. Quasi-universality in single-cell sequencing data. *arXiv preprint arXiv:1810.03602*, 2018.

[ABM18a] Madhu Advani, Guy Bunin, and Pankaj Mehta. Statistical physics of community ecology: a cavity solution to MacArthur's consumer resource model. *Journal of Statistical Mechanics*, 2018:033406, 2018.

[ABM18b] Madhu Advani, Guy Bunin, and Pankaj Mehta. Statistical physics of community ecology: a cavity solution to macarthurÕs consumer resource model. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(3):033406, 2018.

[AF19] Ada Altieri and Silvio Franz. Constraint satisfaction mechanisms for marginal stability and criticality in large ecosystems. *Physical Review E*, 99(1):010401, 2019.

[All20] Stefano Allesina. Going big. *Unsolved Problems in Ecology*, page 374, 2020.

[ASG+16] Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. *Reviews of Modern Physics*, 88(3):035003, 2016.

[AT12] Stefano Allesina and Si Tang. Stability criteria for complex ecosystems. *Nature*, 483(7388):205–208, 2012.

[AT15] Stefano Allesina and Si Tang. The stability–complexity relationship at age 40: a random matrix perspective. *Population Ecology*, 57(1):63–75, 2015.

[AVDB18] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[BA17] Matthieu Barbier and Jean-François Arnoldi. The cavity method for community ecology. *bioRxiv*, page 147728, 2017.

[BABL18] Matthieu Barbier, Jean-François Arnoldi, Guy Bunin, and Michel Loreau. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences*, 115(9):2156–2161, 2018.

[BBC18a] Giulio Biroli, Guy Bunin, and C. Cammarota. Marginally stable equilibria in critical ecosystems. *New Journal of Physics*, 20:083051, 2018.

[BBC18b] Giulio Biroli, Guy Bunin, and Chiara Cammarota. Marginally stable equilibria in critical ecosystems. *New Journal of Physics*, 20(8):083051, 2018.

[Ber99] Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1999.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[BO18] Stacey Butler and James P O'Dwyer. Stability criteria for complex microbial communities. *Nature communications*, 9(1):1–10, 2018.

[BS06] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[Bun17] Guy Bunin. Ecological communities with lotka-volterra dynamics. *Physical Review E*, 95(4):042414, 2017.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[CD11] Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[CFF+18] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[Che90a] Peter Chesson. MacArthur's consumer-resource model. *Theoretical Population Biology*, 37:26, 1990.

[Che90b] Peter Chesson. Macarthur's consumer-resource model. *Theoretical Population Biology*, 37(1):26–38, 1990.

[Che00] Peter Chesson. Mechanisms of maintenance of species diversity. *Annual review of Ecology and Systematics*, 31:343, 2000.

[CHM+18] Paul I Costea, Falk Hildebrand, Arumugam Manimozhiyan, Fredrik Bäckhed, Martin J Blaser, Frederic D Bushman, Willem M De Vos, S Dusko Ehrlich, Claire M Fraser, Masahira Hattori, et al. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1):8, 2018.

[CL03] Jonathan M Chase and Mathew A Leibold. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, Chicago, IL, 2003.

[CLW+11] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the national academy of sciences*, 108(Supplement 1):4516–4522, 2011.

[CMIM19] Wenping Cui, Robert Marsland III, and Pankaj Mehta. Diverse communities behave like typical random ecosystems. page arXiv:1904.02610, 2019.

[CMIM20] Wenping Cui, Robert Marsland III, and Pankaj Mehta. Effect of resource dynamics on species packing in diverse ecosystems. *Physical Review Letters*, 125(4):048101, 2020.

[CRM20] Wenping Cui, Jason W Rocks, and Pankaj Mehta. The perturbative resolvent method: spectral densities of random matrix ensembles via perturbation theory. *arXiv preprint arXiv:2012.00663*, 2020.

[dAT78] Jairo RL de Almeida and David J Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.

[DB20] Itay Dalmedigos and Guy Bunin. Dynamical persistence in resource-consumer models. *arXiv preprint arXiv:2002.04358*, 2020.

[DFM16] Benjamin Dickens, Charles K Fisher, and Pankaj Mehta. Analytically tractable model for community ecology with many species. *Physical Review E*, 94(2):022423, 2016.

[DS07] R Brent Dozier and Jack W Silverstein. On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, 98(4):678–694, 2007.

[DVR+18] Michaël Dougoud, Laura Vinckenbosch, Rudolf P Rohr, Louis-Félix Bersier, and Christian Mazza. The feasibility of equilibria in large ecosystems: A primary but neglected concept in the complexity-stability debate. *PLoS computational biology*, 14(2):e1005988, 2018.

[Dys62] Freeman J Dyson. Statistical theory of the energy levels of complex systems. i. *Journal of Mathematical Physics*, 3(1):140–156, 1962.

[FHG17] Jonathan Friedman, Logan M Higgins, and Jeff Gore. Community structure follows simple assembly rules in microbial microcosms. *Nature ecology & evolution*, 1:0109, 2017.

[FM14] Charles K Fisher and Pankaj Mehta. The transition between the niche and neutral regimes in ecology. *Proceedings of the National Academy of Sciences*, 111(36):13111–13116, 2014.

[FMW17] Charles K Fisher, Thierry Mora, and Aleksandra M Walczak. Variable habitat conditions drive species covariation in the human microbiota. *PLoS computational biology*, 13(4):e1005435, 2017.

[FP16] Silvio Franz and Giorgio Parisi. The simplest model of jamming. *Journal of Physics A: Mathematical and Theoretical*, 49(14):145001, 2016.

[GAS+17] Jacopo Grilli, Matteo Adorisio, Samir Suweis, György Barabás, Jayanth R Banavar, Stefano Allesina, and Amos Maritan. Feasibility and coexistence of large ecological communities. *Nature communications*, 8:14389, 2017.

[GBMSA17] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[GGRA18] Theo Gibbs, Jacopo Grilli, Tim Rogers, and Stefano Allesina. Effect of population abundances on the stability of large random ecosystems. *Physical Review E*, 98(2):022410, 2018.

[GI83] Donald Goldfarb and Ashok Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical programming*, 27(1):1–33, 1983.

[Gin65] Jean Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449, 1965.

[GLB+18] Joshua E. Goldford, Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361:469, 2018.

[GRA16] Jacopo Grilli, Tim Rogers, and Stefano Allesina. Modularity and stability in ecological communities. *Nature communications*, 7(1):1–10, 2016.

[GS10] Daniel Gourion and Alberto Seeger. Deterministic and stochastic methods for computing volumetric moduli of convex cones. *Computational & Applied Mathematics*, 29(2):215–246, 2010.

[HGK+12a] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H Badger, Asif T Chinwalla, Heather H Creasy, Ashlee M Earl, Michael G FitzGerald, Robert S Fulton, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207, 2012.

[HGK+12b] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H Badger, Asif T Chinwalla, Heather H Creasy, Ashlee M Earl, Michael G FitzGerald, Robert S Fulton, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207, 2012.

[HJ13] Tori M Hoehler and Bo Barker Jørgensen. Microbial life under extreme energy limitation. *Nature Reviews Microbiology*, 11(2):83, 2013.

[HRD+14] William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, Christopher J. Marx, and Daniel Segrè. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7:1104, 2014.

[Isi25] Ernst Ising. Contribution to the theory of ferromagnetism. *Z. Phys.*, 31:253–258, 1925.

[KB12] Eugenius Kaszkurewicz and Amit Bhaya. *Matrix diagonal stability in systems and computation*. Springer Science & Business Media, 2012.

[KMS+12] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[KS15] David A Kessler and Nadav M Shnerb. Generalized model of island biodiversity. *Physical Review E*, 91(4):042705, 2015.

[LE20] Stefan Landmann and Andreas Engel. Large systems of random linear equations with nonnegative solutions: Characterizing the solvable and the unsolvable phase. *Physical Review E*, 101(6):062119, 2020.

[Lei95] Matthew A Leibold. The niche concept revisited: mechanistic models and community context. *Ecology*, 76(5):1371–1382, 1995.

[LLL+19] Zhiyuan Li, Bo Liu, Sophia Hsin-Jung Li, Christopher G King, Zemer Gitai, and Ned S Wingreen. Modeling microbial diversity with metabolic trade-offs. *bioRxiv*, page 664698, 2019.

[LN10] Andrea J Liu and Sidney R Nagel. The jamming transition and the marginally jammed solid. *Annu. Rev. Condens. Matter Phys.*, 1(1):347–369, 2010.

[LNV18] Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. *Introduction to random matrices: theory and practice*, volume 26. Springer, 2018.

[LV+11] Philippe Loubaton, Pascal Vallet, et al. Almost sure localization of the eigenvalues in a gaussian information plus noise model—application to the spiked models. *Electron. J. Probab*, 16:1934–1959, 2011.

[M+06] George Marsaglia et al. Ratios of normal variables. *Journal of Statistical Software*, 16(4):1–10, 2006.

[MA77] Richard McGehee and Robert A. Armstrong. Some Mathematical Problems Concerning the Ecological Principle of Competitive Exclusion. *Journal of Differential Equations*, 23:30, 1977.

[Ma18] Shang-Keng Ma. *Modern theory of critical phenomena*. Routledge, 2018.

[Mac70] Robert MacArthur. Species packing and competitive equilibrium for many species. *Theoretical population biology*, 1(1):1–11, 1970.

[May72] Robert M May. Will a large complex system be stable? *Nature*, 238(5364):413, 1972.

[MBW+19] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre GR Day, Clint Richardson, Charles K Fisher, and David J Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, (in press), 2019.

[McC00] Kevin Shear McCann. The diversity–stability debate. *Nature*, 405(6783):228–233, 2000.

[MCGM20] Robert Marsland, Wenping Cui, Joshua Goldford, and Pankaj Mehta. The community simulator: A python package for microbial ecology. *Plos one*, 15(3):e0230430, 2020.

[MCM20] Robert Marsland, Wenping Cui, and Pankaj Mehta. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Scientific reports*, 10(1):1–17, 2020.

[MCWMI19] Pankaj Mehta, Wenping Cui, Ching-Hao Wang, and Robert Marsland III. Constrained optimization as ecological dynamics with applications to random quadratic programming in high dimensions. *Physical Review E*, 99(5):052111, 2019.

[MICG+19] Robert Marsland III, Wenping Cui, Joshua Goldford, Alvaro Sanchez, Kirill Korolev, and Pankaj Mehta. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLOS Computational Biology*, 15(2):e1006793, 2019.

[MICM20] Robert Marsland III, Wenping Cui, and Pankaj Mehta. The minimum environmental perturbation principle: A new perspective on niche theory. *The American Naturalist*, 196(3):291–305, 2020.

[ML67a] Robert Macarthur and Richard Levins. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921):377–385, 1967.

[ML67b] Robert Macarthur and Richard Levins. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921):377–385, September 1967.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MM11] Cristopher Moore and Stephan Mertens. *The nature of computation*. OUP Oxford, 2011.

[MP67a] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[MP67b] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[MP03] Marc Mézard and Giorgio Parisi. The cavity method at zero temperature. *Journal of Statistical Physics*, 111(1-2):1–34, 2003.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[MPZ02] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[Mur07] James D Murray. *Mathematical biology: I. An introduction*, volume 17. Springer Science & Business Media, 2007.

[PGB11] Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

[PTW17] Anna Posfai, Thibaud Taillefumier, and Ned S Wingreen. Metabolic trade-offs promote diversity in a model ecosystem. *Physical Review Letters*, 118(2):028103, 2017.

[QLR+10] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59, 2010.

[RBW+13] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[RCKT08] Tim Rogers, Isaac Pérez Castillo, Reimer Kühn, and Koujin Takeda. Cavity approach to the spectral density of sparse symmetric random matrices. *Physical Review E*, 78(3):031116, 2008.

[Ric57] Bellman Richard. Dynamic programming. *Princeton University Press*, 89:92, 1957.

[RMS15] Mohammad Ramezanali, Partha P Mitra, and Anirvan M Sengupta. The cavity method for analysis of large-scale penalized regression. *arXiv preprint arXiv:1501.03194*, 2015.

[RSB14] Rudolf P Rohr, Serguei Saavedra, and Jordi Bascompte. On the structural stability of mutualistic systems. *Science*, 345(6195):1253497, 2014.

[SCC+15] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015.

[SCG+18] Carlos A Serván, José A Capitán, Jacopo Grilli, Kent E Morrison, and Stefano Allesina. Coexistence of many species in random ecosystems. *Nature ecology & evolution*, 2(8):1237–1242, 2018.

[SCS88] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[SWGV14] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[Til82a] David Tilman. *Resource competition and community structure*, volume 17. Princeton University Press, 1982.

[Til82b] David Tilman. *Resource Competition and Community Structure.* Princeton University Press, Princeton, NJ, 1982.

[TM17] Mikhail Tikhonov and Remi Monasson. Collective phase in resource competition in a highly diverse ecosystem. *Physical Review Letters*, 118(4):048103, 2017.

[TPMW17] Thibaud Taillefumier, Anna Posfai, Yigal Meir, and Ned S. Wingreen. Microbial consortia at steady supply. *eLife*, 6:e22644, 2017.

[TSM+17] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset,

Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, and Earth Microbiome Project Consortium. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551:457, 2017.

[WAP+16] Stefanie Widder, Rosalind J Allen, Thomas Pfeiffer, Thomas P Curtis, Carsten Wiuf, William T Sloan, Otto X Cordero, Sam P Brown, Babak Momeni, Wenying Shou, Helen Kettle, Harry J Flint, Andreas F Haas, Béatrice Laroche, Jan-Ulrich Kreft, Tobias Großkopf, Jef Huisman, Andrew Free, Cristian Picioreanu, Christopher Quince, Isaac Klapper, Simon Labarthe, Barth F. Smets, Harris Wang, and Orkun S Soyer. Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME journal*, 10:2557, 2016.

[Wig58] Eugene P Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pages 325–327, 1958.

[WT03] Max Welling and Yee Whye Teh. Approximate inference in boltzmann machines. *Artificial Intelligence*, 143(1):19–50, 2003.

[YFW01] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Generalized belief propagation. In *Advances in neural information processing systems*, pages 689–695, 2001.

[YSL+19] Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S Marks, John Ingraham, and Chris Sander. Interpretable machine learning for perturbation biology. *bioRxiv*, page 746842, 2019.

[Zde09] Lenka Zdeborová. Statistical physics of hard optimization problems. *Acta Physica Slovaca. Reviews and Tutorials*, 59(3):169–303, 2009.

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[ZS16] Ali R. Zomorrodi and Daniel Segrè. Synthetic ecology of microbes: Mathematical models and applications. *J. Mol. Biol.*, 428:837, 2016.