UNIVERSALITY IN MULTIPARAMETER FITTING:

SLOPPY MODELS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Joshua James Waterfall

September 2006

UNIVERSALITY IN MULTIPARAMETER FITTING: SLOPPY MODELS

Joshua James Waterfall, Ph.D.

Cornell University 2006

In order to understand a variety of physical phenomena (such as signaling networks in molecular biology or crystal structures in condensed matter physics), scientists often develop models with many unknown or tunable parameters. Such multi-parameter models and systems are often *sloppy*. For practical purposes their behavior depends only on a few stiffly constrained combinations of parameters; other directions in parameter space can change by orders of magnitude without significantly changing the behavior. We develop the theoretical basis of sloppiness and argue that there is in fact a new universality class to which these models belong.

We begin by defining sloppiness (an exponentially large range of sensitivity to different combinations of parameters, with a roughly uniform distribution of sensitivities between the extremes). We then document sloppiness in a variety of models from different scientific fields. Several mathematically well-defined classes of models, some sloppy and some not sloppy, are then analyzed to understand the origins of sloppiness. Drawing connections to the field of random matrix theory, we derive an ensemble of sloppy models. The heart of sloppiness in this ensemble is shown to be the Vandermonde matrix. By demonstrating the novel statistical properties of this ensemble we argue that it constitutes a new universality class. Inspired by the properties of this Vandermonde ensemble we develop new tools for

analyzing complex, real-world models with many parameters.

In the final section we focus on a particular complex, real-world model with many parameters. We formulate and analyze a mathematical description of the quorum sensing network in the bacterium *Agrobacterium tumefaciens*. This network allows *Agrobacterium* to regulate gene expression in accordance with its population density. The mathematical description includes twenty four unknown parameters quantifying the biochemical interactions. While not complete, the model provides insight into the quorum sensing process and we suggest ways of coupling the model with experiments in the future.

BIOGRAPHICAL SKETCH

The author was born in Fairfax, Virginia on May 25, 1979 to Milde and James Waterfall and with the exception of nine months in Alaska, the author was raised in northern Virginia. After attending the Thomas Jefferson High School for Science and Technology in Alexandria, Virginia the author enrolled at the College of William and Mary in Williamsburg, Virginia. The author majored in both Physics as well as Applied Mathematics, becoming the first Applied Math major in the three hundred eight year history of the College. Highlights of these undergraduate years include performing research under Dr. Mei-Yin Chou at the Georgia Institute of Technology in the summer of 2000, conducting a senior thesis under Dr. Dennis Manos for which he was awarded highest honors, and being one of the final members of the Sigma Nu fraternity. Upon graduation in 2001 with a Bachelor of Science and membership in the Phi Beta Kappa honor society (which was founded at William and Mary in 1776), the author immediately enrolled in graduate school in Physics at Cornell University. Quite quickly the author joined the research group of Dr. James Sethna studying biological signaling networks through computational modeling with a number of exciting collaborations. To pursue those interests the author was awarded a Computational Science Graduate Fellowship from the Department of Energy. Highlights of these years include a summer working with Dr. Daniel Rokhsar at the Joint Genome Institute and getting married to Heidi Elston on July 2, 2005. The author will be continuing his stay at Cornell and deepening his immersion in molecular biology upon graduation, enjoying a postdoctoral position with John Lis in the Department of Molecular Biology and Genetics.

To my parents.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Origins of Sloppiness

1.1   Introduction

In a variety of contexts, physicists and other scientists study complex, nonlinear
models with many unknown or tunable parameters to explain experimental data
and predict future experiments. In Figure 1.1 we see (a) a model of a biological
signaling network, (b) its fit to previously collected data (for example the time
course of the active state of the protein Erk in response to different growth factors),
and (c) its prediction of a future experiment (the Erk activity levels while one of the
proteins in the network, PI3K, is inhibited). In Figure 1.2 we see an even simpler
system—fitting the exponents in a sum of exponentials to data from radioactive
decay in order to determine the elements in the sample and to predict the future
decay time course. In both cases we have a model with free parameters and we
have a set of data. We quantify the difference between the data and the model for
a given set of parameters by a suitable cost function (e.g. χ^2 or log-likelihood),
and then we study the dependence of the model behavior on the parameter values
by studying how the cost rises away from the best fit.
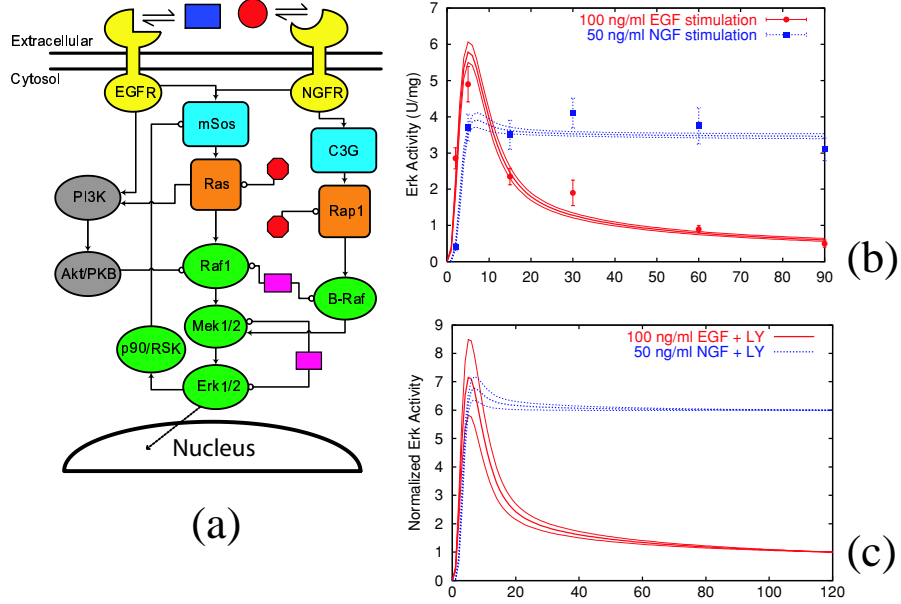
Figure 1.1: Fitting and predicting with multiparameter models of biological networks. A model (a) of the protein interactions defining growth factor signaling in PC12 cells. The mathematical description of this model (coupled first-order nonlinear ordinary differential equations) contains 48 free parameters (rate and Michaelis-Menten constants) that can be fit to previously collected data, for example (b), the time course of Erk activity in response to two different growth factors. The model can then be used to predict future experiments, such as (c) Erk activity while the activity of the protein PI3K (topmost grey oval in (a)) is inhibited. In (b) and (c) the horizontal axes are time in minutes. Figures are from reference [4].

(a)   (b)
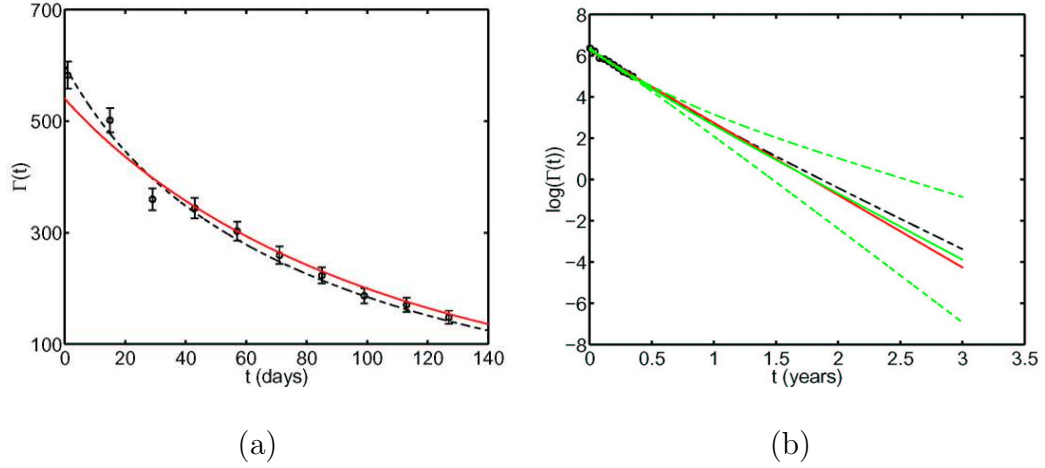
Figure 1.2:   Fitting and predicting with multiparameter models of radioactive decay. The model describing this decay is a sum of three exponentials where both the three decay rates and the three intial amounts are unknown parameters. In (a) we see a set of data points with error bars, the 'true' exponential describing the net decay as a dashed black line, and one of a number of good fits from the model as a solid red line. In (b) we use the model to predict the radiation at late times. The datapoints visible in the upper left, the dashed black line, and the solid red line are the same as in (a). The best fit is not tightly constrained and the solid green line is the prediction from an ensemble of parameter sets, each weighted by their fit to the data. The dashed green lines are the standard deviation of this prediction over the ensemble of parameter sets.

We explain why such systems so often are *sloppy*; for practical purposes their behavior depends only on a few stiffly constrained combinations of the parameters; other directions in parameter space can change by orders of magnitude without significantly changing the behavior. We contrast examples of sloppy models (from systems biology, variational quantum Monte Carlo, and common data fitting) with systems which are not sloppy (multidimensional linear regression, random matrix ensembles). We observe that the eigenvalue spectra for the sensitivity of sloppy models have a striking, characteristic form, with a density of logarithms of eigenvalues which is roughly constant over a large range. We suggest that the common features of sloppy models indicate that they may belong to a common universality class. In particular, we motivate focusing on a *Vandermonde ensemble* of multiparameter nonlinear models and show in one limit that they exhibit the universal features of sloppy models.

Given a suitable cost function $C(\mathbf{p})$ measuring the change in system behavior as the parameters $\mathbf{p}$ vary from their original values $\mathbf{p}^{(0)}$ (e.g., a sum of squared residuals), we are interested in the shape of the cost function landscape. Figure 1.3 contrasts a stiff and sloppy direction for the dependence of the radioactivity of a mixture of radionuclides on their decay lifetimes. One must change parameters along the sloppy direction over a thousand times more than along the stiff direction in order to change the behavior by the same amount.

The stiff and sloppy directions can be quantified as eigenvalues and eigenvectors of the Hessian of the cost:

$$H_{ij} = \frac{\partial^2 C}{\partial p_i \partial p_j}\bigg|_{\mathbf{p}^{(0)}}. \tag{1.1}$$

The Hessian tells us the curvature of the cost function in the neighborhood of the point $\mathbf{p}^{(0)}$, approximating the fully nonlinear (bumpy, windy) surface by an ellipti-

cal bowl. The eigenvectors (linear combinations of the original, bare, parameters) of the Hessian are the principle axes of this ellipse and the square root of the corresponding eigenvalue is the curvature is along that eigendirection. The horizontal and vertical directions in Figure 1.3 are eigenvectors of that particular model. Figure 1.4 shows the eigenvalues of the cost Hessian for many different systems; those in (a), (b), (c), (d) and (h) are all sloppy. Since the sensitivity of model behavior to changes along an eigenvector is given by the square root of the eigenvalue, the range in eigenvalues of roughly one million for the sloppy models means their cost-contours have aspect ratios of one thousand, just as in Figure 1.3. Although anharmonic effects rapidly become important along sloppy directions, as can be seen in Figure 1.3, a principal component analysis of a Monte-Carlo sampling of low-cost states has a similar spectrum of eigenvalues [5]; the sloppy eigendirections become curved sloppy manifolds in parameter space. Similar sloppy behavior has been demonstrated in fourteen systems biology models taken from the literature [4, 16], and in three multiparameter interatomic potentials fit to electronic structure data [30]. In these disparate models we see a common, peculiar behavior: the nth stiffest eigendirection is more important than the $(n+1)$th by a roughly constant factor, giving a total range of eigenvalues of typically over a million for any model with more than eight parameters. We call systems exhibiting these characteristic features *sloppy models*.

This sloppiness has a number of important implications. In estimating prediction errors, sloppiness affects both the estimation of statistical errors due to uncertainties in the experimental data [4, 16] and allows an estimation of systematic errors due to imperfections in the models (for example in interatomic potentials [30] and density functional theory [23]). It makes extracting parameter values

5
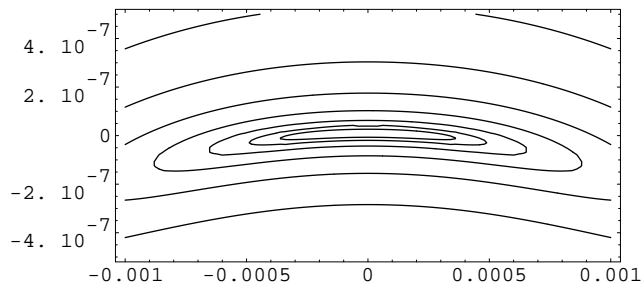
Figure 1.3: Cost contours for fitting exponentials. Here we show contours of constant cost $C(\boldsymbol{\gamma})$ for the radioactive activity of a mixture of twelve common radionuclides, as a function of relative changes in their decay constants $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{12})$. (The radionuclides chosen are those available from Perkin-Elmer [27] with half-lives less than 100 days.) The plot shows a cross-section along the eigendirections corresponding to the second-stiffest eigenvalue (vertical) and the sloppiest eigenvalue (horizontal). Note that the horizontal axis has been compressed by a factor of one thousand; the aspect ratio is actually comparable to a one-inch human hair. The sloppiness is not just an artifact of the harmonic approximation. Although anharmonic effects rapidly become important along the sloppy eigendirections as shown here, a principle-component analysis of a Monte-Carlo sampling of low-cost states has a similar spectrum of eigenvalues [5]; the sloppy eigendirections become curved sloppy manifolds in parameter space.
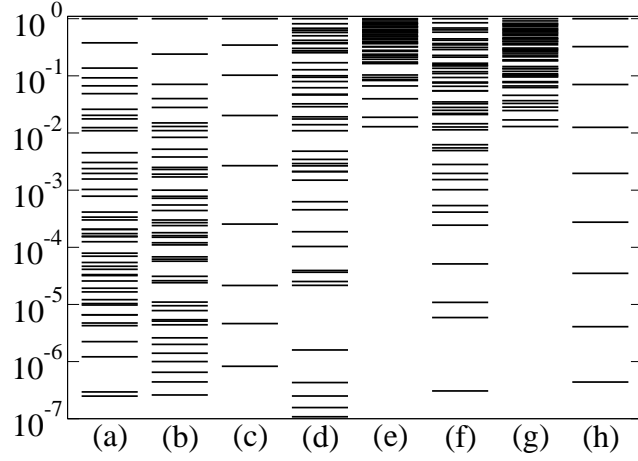
Figure 1.4: Hessian eigenvalues of various multiparameter models. Eigenvalues giving the stiffness/sloppiness of various models as parameters are varied. Each spectrum has been shifted so that the largest eigenvalue is one. (a) Growth factor signaling model (coupled nonlinear ODEs) for PC12 cells [4], as the 48 parameters (rate and Michaels-Menten constants) are varied. (b) Variational wave-function used in quantum Monte-Carlo, as the Jastrow parameters (for electron-electron coincidence cusps) are varied, (c) Radioactivity time evolution for a mixture of twelve common radionuclides as the half-lives γ_i are varied. The radionuclides are those available from Perkin-Elmer [27] with half-lives less than 100 days. (Only the first nine eigenvalues are shown.) (d) The same exponential decay model as in (c) with 48 decay constants γ_i randomly spread over a range of e^{50}. (e) One random 48×48 matrix in the Gaussian Orthogonal Ensemble (GOE) (not sloppy). (f) A product of five random 48×48 matrices, illustrating the random product ensemble (not sloppy, but ill conditioned). (g) A plane in 48 dimensions fit to 68 data points, the same number and data points as for the biology model in column (a) (Wishart statistics, not sloppy). (h) A polynomial fit to data, as the 48 monomial coefficients are varied (the Hilbert matrix [17], sloppy).

from fits to sloppy models ill-posed [4, 13]. Conversely, it is much more efficient to improve the predictivity of a model by fitting parameters to system behavior than by designing experiments that precisely determine the individual parameter values [16]. Sloppy problems are also better approached with optimization algorithms [28, 7] (like the Levenberg–Marquardt and Nelder–Mead methods) which can adapt to widely diverging step sizes along different parameter combinations.

In this thesis we focus on explaining why sloppy behavior arises and where it can be expected to manifest itself. We first explain in more detail the 'real-life' sloppy models; the biological network of Figure 1.4 (a) is detailed in Section 1.2 and the quantum mechanical wave function of Figure 1.4 (b) is described in Section 1.3. In Section 1.4 we examine sloppiness in the problem of fitting sums of exponentials, as in Figure 1.4 (c) and (d). We then turn to transforming between sloppy and unsloppy parameterizations in Section 1.5, illustrating the process with fitting polynomials as in Figure 1.4 (h). Section 1.6 expounds on Figure 1.4 (g) by analyzing classical multiple linear regression models and their corresponding Wishart statistics, demonstrating that they are not sloppy. In the following section, 1.7, we also contrast our sloppy models with the ensembles of Random Matrix Theory (Figure 1.4 (e) and (f)), finding that they do not describe sloppiness either. In Section 1.8 we suggest that there is a universality class of sloppy models, and we analyze a particular ensemble of models to give an analytic explanation for their sloppy behavior. Lastly, Section 1.9 details the effects of coupling multiple models from this ensemble and the connections to 'real-life' sloppy models.

1.2 Biological Networks

The growth factor signaling model in Figure 1.4 (a) is depicted in Figure 1.5 [4, 5, 16]. It describes the network of interactions by which PC12 cells (a rat adrenal pheochromocytoma cell line) either differentiate or proliferate in response to growth factor signals. The network begins with the extracellular concentration of two growth factors, neuronal growth factor (NGF) and epidermal growth factor (EGF) and ends with the activation (phosporylation) state of the Erk protein. In real cells active Erk then translocates into the nucleus and controls gene expression but this is not included in the model. The model consists of the concentration of thirty two chemical species (i.e. peptides and proteins in various modified forms or complexes) coupled by a system of twenty eight first order nonlinear ordinary differential equations describing the biochemical reactions that constitute the network. These equations include forty eight unknown parameters (rate and Michaelis-Menten constants) that were fit to sixty eight data points from fourteen cell biology experiments (time series Western blots with either wild-type or single transfections and various initial concentrations of growth factors).

The sloppiness of this model is not unique in the field of biological network modeling. Characteristically sloppy sensitivity spectra have been identified in an array of models ranging from the yeast cell cycle to circadian rhythms in *Drosophila* to neurotransmitter signaling in humans [16].

1.3 Quantum Monte Carlo

As a variational description of the various eigenstates of the Hamiltonian, quantum mechanical many-body wavefunctions are used to calculate the electronic structure
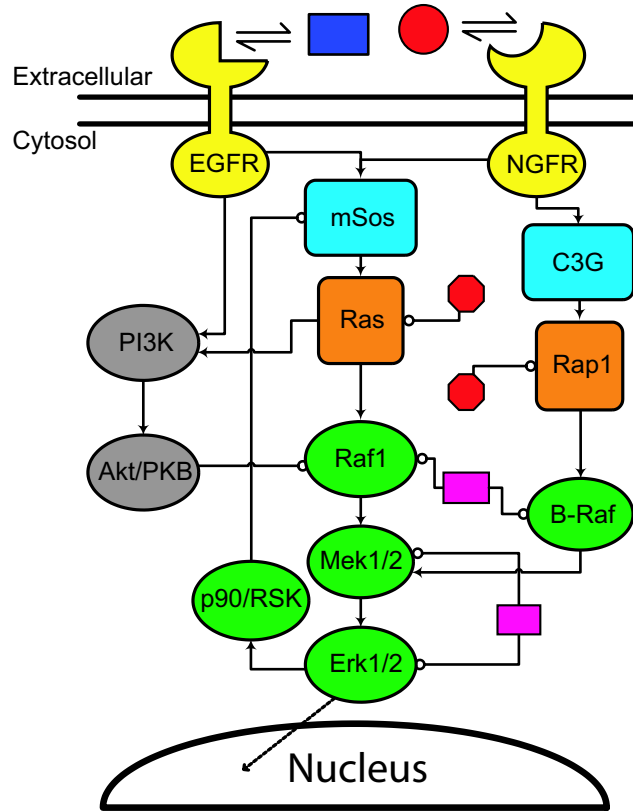
Figure 1.5: Model of growth factor signaling in PC12 cells [4]. The blue box is epidermal growth factor (EGF) and the red circle is neuronal growth factor (NGF). The red octagons are GTPase Activating Proteins (GAPs) that inhibit the signaling activity of the G-proteins Ras and Rap1. The purple boxes are phosphatases that inhibit the signaling activity of the kinases Raf1 and B-Raf.

of atoms and molecules. The wavefunctions are parameterized by two classes of parameters—Configuration Interaction (CI) coefficients for expansions in determinants of single particle orbitals and Jastrow parameters to describe the cusps that occur at electron-electron coincidences [24]. The spectrum in Figure 1.4 (b) is the sensitivity for the Jastrow parameters alone, the CI coefficients were held fixed in the calculation [32].

Given a variational wave function, there are two types of optimization that may be done. The more obvious type is *energy minimization*. The energy of the true ground state of the system is a lower bound on the energy of any possible wavefunction so we may obtain a good approximation of the ground state by minimizing the energy of our trial wavefunction. The second type of variational optimization, *variance minimization*, is both more subtle and more powerful. This approach is based on the fact that for any true eigenstate of the Hamiltonian, the variance of any observable that commutes with the Hamiltonian must be zero. Since all variances must be strictly positive, any wave function that is not an eigenstate must have a higher variance. Therefore one can obtain approximations to eigenstates by adjusting parameters in a trial wavefunction to minimize such a variance. This method is superior to standard energy minimization for several reasons: (1) the convergence of the minimization calculations is considerably easier to validate since variances are bounded below by zero while the (ground state) lower bound on energy calculations is either unknown *a priori* or nonexistent (relativistic Hamiltonians are unbounded below), (2) since the variance is a sum-of-squares function, sophisticated optimization algorithms (e.g. Levenberg-Marquardt) can be used to exploit this structure efficiently, (3) the zero-variance principle holds for any eigenstate, not simply the ground state, so approximations to any excited state

can be obtained with this approach. The minimization that lead to Figure 1.4 (b) was just such a variance minimization [32].

1.4   Exponentials

Fitting decay constants to data that is a sum of exponentials is a famously ill-posed problem [20, 33]. Consider a mixture of equal amounts of N radioactive elements, whose decay signal is thus the sum of N exponentials with decay rates $\boldsymbol{\gamma}^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_N^{(0)})$:

$$y(t, \boldsymbol{\gamma}^{(0)}) = \sum_{i=1}^{N} \exp(-\gamma_i^{(0)} t). \tag{1.2}$$

Unless the individual lifetimes are well separated, the net radiation cannot be used to measure the lifetimes reliably. The difficulty is that the signal is the sum of many functions with similar shapes; one can generate almost identical signals with wildly different values for the parameters. We define a cost function by integrating the square of the difference between $y(t, \boldsymbol{\gamma})$ and $y(t, \boldsymbol{\gamma}^{(0)})$ over $\mathrm{d}\log t = \mathrm{d}t/t$:

$$r(t, \boldsymbol{\gamma}) = \sum_{i} \exp(-\gamma_i t) - \sum_{i} \exp(-\gamma_i^0 t) \tag{1.3}$$

$$C(t; \boldsymbol{\gamma}) = \int_o^\infty r^2(t, \boldsymbol{\gamma}) \frac{\mathrm{d}t}{t}. \tag{1.4}$$

Spacing the 'data points' equally in logarithmic time makes analyzing large ranges of decay constants γ convenient. If the data were spaced evenly in regular time, the slow decay rates become much more significant in describing the data for the trivial reason that they are given too much time as $t \to \infty$ to dominate the behavior.

Because the decay constants are positive and can have a large range of sizes, we use their logarithms as our parameters ($p_i = \log \gamma_i$), giving model sensitivity

to relative changes in the decay rates. The resulting Hessian is

$$H_{ij} = \frac{\partial^2 C(\boldsymbol{\gamma})}{\partial \log \gamma_i \partial \log \gamma_j} \tag{1.5}$$

$$= 2 \int_0^\infty \left[\frac{\partial r(t, \boldsymbol{\gamma})}{\partial \log \gamma_i} \frac{\partial r(t, \boldsymbol{\gamma})}{\partial \log \gamma_j} + r(t, \boldsymbol{\gamma}) \frac{\partial^2 r(t, \boldsymbol{\gamma})}{\partial \log \gamma_i \partial \log \gamma_j} \right] \frac{\mathrm{d}t}{t}. \tag{1.6}$$

At the correct parameter values, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$, each residual is zero and the second term in the Hessian drops out. We thus have

$$H_{ij} = 2\gamma_i \gamma_j \int_0^\infty t \exp(-(\gamma_i + \gamma_j)t) \mathrm{d}t. \tag{1.7}$$

Integrating by parts we obtain

$$H_{ij} = 2 \frac{\gamma_i \gamma_j}{(\gamma_i + \gamma_j)^2}. \tag{1.8}$$

For the twelve radionuclides described in the caption to Figure 1.4 (c) and listed in Table 1.1, the eigenvalues of the Hessian are each separated by nearly one decade; the sloppiest mode has an eigenvalue a factor of 10^{10} smaller (less important) than the stiffest.

The sloppiness in sums of exponentials is due to the compensation that can occur between decay rates that are within a decade or so of one another, just as the sloppiness in more complex models is presumably due to the compensation of subsets of parameters with similar effects. The range of eigenvalues for the twelve radioactive decay elements is far larger than that for the 'real-life' systems biology (Figure 1.4 (a)) and variational wavefunction models (Figure 1.4 (b)), and the eigenvalue spacings are much more rigid (a phenomenon called 'level repulsion' that we analyze below). We shall understand both of these effects in detail using the conclusions below; the fitting exponentials problem turns out to be a subset of a large *Vandermonde ensemble* for which the conclusions apply. There we shall see that the large range and rigidly equal spacings are a reflection of the

13

Table 1.1: Perkin-Elmer radionuclide decay rates. Only elements with half-lives under one hundred days were included in analysis [27].

Element	Half-life (days)
Chromium - 51	27.7
Indium - 111	2.83
Iodine - 125	60.14
Iodine - 131	8.04
Iron - 59	44.6
Lutetium - 177	6.71
Phosphorous - 32	14.29
Phosphorous - 33	25.4
Rubidium - 86	18.66
Scandium - 46	83.83
Sulfur - 35	87.4
Yttrium - 90	2.67

relatively narrow range of lifetimes in the twelve elements, which vary over a range of roughly largest/smallest $= 33$ ($p_i = \log \gamma_i$ in a range $2\epsilon \approx 3.5$). If we pick 48 lifetimes whose logarithms are instead uniformly distributed over a range of $2\epsilon = 50$ (largest/smallest $\approx 10^{21}$), the density of levels and the variations in spacings between neighboring levels in the new spectrum (Figure 1.4 (d)) is similar to that of the real-life models in (a) and (b). With this larger range of decay rates, the individual parameters cannot all compensate for one another; the very large decay rates can only exchange with one another and similarly for the very small rates.

Another way to explore the effects of coupling distinct subsets of parameters in this model is to allow the initial concentrations, A_i, to be unknown parameters. The Hessian then has a block structure corresponding to derivatives with respect to these two different classes of parameters.

$$H_{ij} = \begin{pmatrix} \frac{\partial^2 C}{\partial \log A_i \partial \log A_j} & \frac{\partial^2 C}{\partial \log \gamma_i \partial \log A_j} \\ \frac{\partial^2 C}{\partial \log A_i \partial \log \gamma_j} & \frac{\partial^2 C}{\partial \log \gamma_i \partial \log \gamma_j} \end{pmatrix} \qquad (1.9)$$

Calculations similar to those above show that the mixed derivative is given by

$$\frac{\partial^2 C}{\partial \log A_i \partial \log \gamma_j} = \frac{-2 A_i A_j \gamma_j}{\gamma_i + \gamma_j}. \qquad (1.10)$$

The value of the other derivative, $\partial^2 C / \partial \log A_i \partial \log A_j$ is slightly trickier because the integral $\int_o^\infty \exp(-(\gamma_i + \gamma_j)t) \mathrm{d} \log t$ does not converge. Figure 1.6 illustrates the problem—any change in parameters which alters the total initial decay rate (the sum of the initial amounts) has an integral that diverges because there is an infinite amount of logarithmic time before the first decay occurs. In order to avoid this problem we simply remove a degree of freedom from our model by declaring the sum of the initial amounts, A_{total}, to be constant. In this formulation the

15
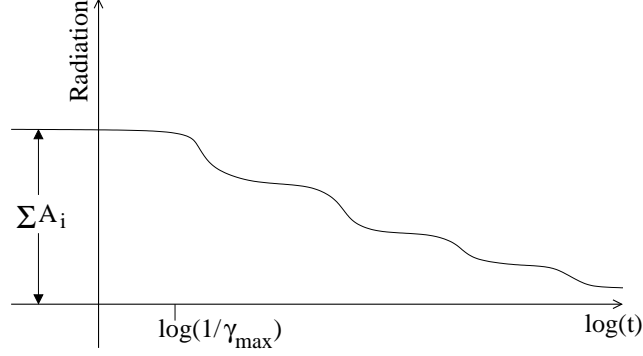
Figure 1.6: A schematic of the discrete nature of radioactive decay. Prior to the first decay (at time $t \approx 1/\gamma_{max}$), the net radiation has a constant level given by the sum of the initial amounts. Since logarithmic time extends to $-\infty$, any change in parameters which alters the sum $\sum_i A_i$ has an infinite cost.

first $N - 1$ initial amounts are free parameters but the final initial amount is not, $A_N \equiv A_{total} - \sum_{i=1}^{N-1} A_i$. The model is then

$$y(t, \boldsymbol{\gamma}, \mathbf{A}) = \sum_{i=1}^{N-1} A_i \exp(-\gamma_i t) + (A_{total} - \sum_{i=1}^{N-1} A_i) \exp(-\gamma_N t). \qquad (1.11)$$

The Hessian with respect to initial amounts is then

$$\frac{\partial^2 C}{\partial \log A_i \partial \log A_j} = A_i A_j \int_0^\infty (\exp(-\gamma_i t) - \exp(-\gamma_N t))(\exp(-\gamma_j t) - \exp(-\gamma_N t))\frac{dt}{t}$$

$$(1.12)$$

$$= A_i A_j \log\left(\frac{(\gamma_i + \gamma_N)(\gamma_j + \gamma_N)}{2\gamma_N(\gamma_i + \gamma_j)}\right). \qquad (1.13)$$

In this new model there are clearly two distinct classes of parameters—the decay rates and the initial amounts. Just as the decay rates can compensate for one another, so too can the initial amounts. Since the two parameter classes affect model behavior in substantially different ways (initial amounts changing the overall level and decay rates changing the characteristic time of the radiation), they can

not compensate for each other. This decomposition of the problem into subsystems which are redundant internally, but not mutually, is significant and we will explore its effects later, when we analyze the statistics of level spacings in the context of the Vandermonde ensemble.

1.5 Polynomials

What makes a model sloppy? We can gain insight by considering the common task of fitting polynomials to data. Whatever the source of the data, if it consists of pairs of points (e.g. one dependent and one independent variable) we can describe the relationship between the variables by a polynomial of some degree. The motivation may be a Taylor series expansion, where the coefficients of the monomials give us the derivatives of some function in a local vicinity, or simply convenience since polynomials are a familiar family of functions we can easily picture.

The first step in fitting this data would be to rescale the dependent variables so that they lie between 0 and 1. Let us call the rescaled dependent variable x. It is a trivial matter to rescale our fit polynomial back to the original range of the data and this formality facilitates analytic results. As elsewhere in this thesis, we define a sum of squared residuals cost function which we wish to minimize, $C(\mathbf{p}) = \sum_i (f(x_i; \mathbf{p}) - y_i)^2$.

We are now faced with a choice that will become important: how do we parameterize our polynomial of degree K, $f(x, \mathbf{p})$? Perhaps the most obvious choice is as a sum of monomials. This is certainly a justifiable set of basis functions — they are familiar, easy to interpret, and may be dictated by the model (e.g. if calculating a Taylor expansion, it is precisely the coefficients of the monomials that we are

(a)   (b)
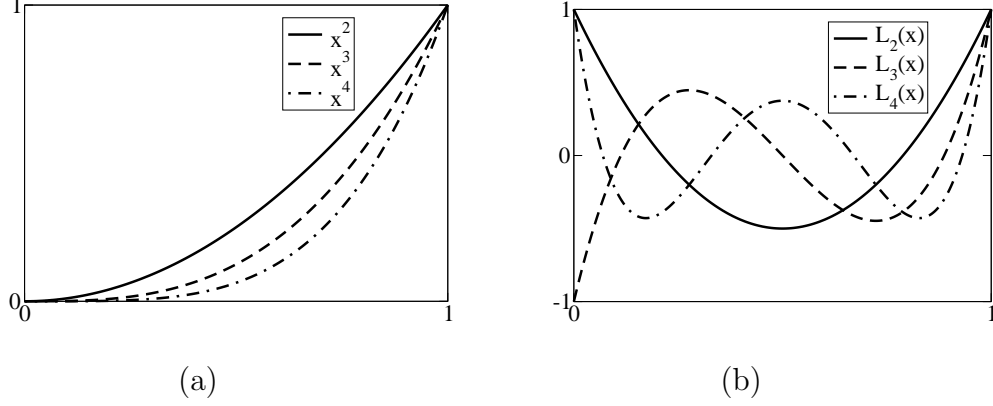
Figure 1.7:   Alternative bases for fitting polynomials. (a) monomials of degree two, three and four, (b) shifted Legendre polynomials of degree two ($L_2(x) = \sqrt{5}(6x^2 - 6x + 1)$), three ($L_3(x) = \sqrt{7}(20x^3 - 30x^2 + 12x - 1)$), and four ($L_4(x) = \sqrt{9}(70x^4 - 140x^3 + 90x^2 - 20x + 1)$). For display purposes the polynomials have been rescaled to have similar maxima and minima.

derive). An alternative choice is to use the coefficients of $f(x)$ expanded as a sum of the shifted Legendre polynomials, $L_k(x), k \in 0 \ldots K$. These polynomials are explicitly defined to be orthonormal on the range [0,1] ($\int_0^1 L_i(x)L_j(x)\mathrm{d}x = \delta_{ij}$) and there is precisely one polynomial of each degree. In Figure 1.7 we depict several polynomials of each type for comparison.

If we choose to use the monomials as our basis, then our function becomes $f(x; \mathbf{p}) = \sum_{k=0}^{K} p_k x^k$. As in the rest of this thesis we are predominantly concerned with the behavior of the true model with the true parameters, so let $\mathbf{p}^{(0)}$ be the true set of monomial coefficients. For this analysis, let us take the limit where the number of data points goes to infinity and the sum becomes an integral. The cost function then is

$$C(\mathbf{p}) = \int_0^1 (\sum_{k=0}^{K} p_k x^k - \sum_{k=0}^{K} p_k^{(0)} x^k)^2 \mathrm{d}x. \tag{1.14}$$

18

The Hessian matrix is

$$H_{ij} = \frac{\partial^2 C(\mathbf{p})}{\partial p_i \partial p_j} = 2 \int_0^1 x^{i+j} \mathrm{d}x = \frac{2}{i+j+1}. \tag{1.15}$$

Aside from the factor of 2, this is the famous Hilbert matrix, A_K [17]. The Hilbert matrix is often cited as a prototypically ill-conditioned matrix and the eigenvalues for the 48×48 Hilbert matrix are shown in Figure 1.4 (h). Indeed, the coefficients of the monomials are known to be poorly determined in such polynomial fits [28].

If we had instead chosen to expand in the shifted Legendre polynomial basis, the Hessian matrix would be

$$H_{ij} = 2 \int_0^1 L_i(x) L_j(x) \mathrm{d}x = 2\delta_{ij}, \tag{1.16}$$

twice the Identity matrix. By changing our parameterization from monomial coefficients to coefficients in the appropriate orthonormal basis, our sloppiness is completely cured. The sloppiness is due to the fact that the monomial coefficients (natural from many perspectives) are a perverse set of coordinates from the point of view of the behavior of the resulting polynomial. We can quantify this by noting that the transformation S_K from the monomial basis to the orthonormal basis (the coefficients of the shifted Legendre polynomials) has a tiny determinant, and therefore the volume enclosed by the monomial basis vectors shrivels and becomes greatly distorted under the transformation. This determinant can be found by noting that

$$H^m = T_{m \to l}^\top H^l T_{m \to l} \tag{1.17}$$

where $H^m = 2A_K$ is the Hessian in the monomial basis, $H^l = 2I$ is the Hessian in the basis of shifted Legendre polynomials, and $T_{m \to l} = S_K$ is the transformation which maps from the monomial basis into the shifted Legendre polynomial basis.

19

Thus $A_K = S_K^\top S_K$ and

$$\det S_K = \sqrt{\det A_K} = \frac{\prod_{i=1}^{K-1}(i!)^2}{\sqrt{\prod_{j=1}^{2K-1}(j!)}} \qquad (1.18)$$

where the last result uses the known determinant of the Hilbert matrix [17]. Since the kth monomial and shifted Legendre polynomial are each of degree k, S_K must be upper triangular and we see that it is in fact the Cholesky decomposition of A_K. Physically, the monomials all have roughly the same shape (starting flat near zero, and rising sharply at the end near one), and can be exchanged for one another, while the orthogonal polynomials each have quite distinct shapes and their contributions to the total model are thus much more identifiable.

In nonlinear sloppy models the sloppiness is more difficult to remove: (a) the transformation to unsloppy parameters will be nonlinear away from the optimum, often not even single-valued, (b) we may not have the insight or the ability to change parameterizations to those natural for fitting purposes, and (c) often the natural parameterization is determined by the science (as in biochemical rate constants, arbitrary linear combinations of which are not biologically motivated).

While this model is pedagogically useful, the fact that every instance of fitting monomials has (twice) the Hilbert matrix as the Hessian and every instance of fitting shifted Legendre polynomials has (twice) the Identity matrix as the Hessian is a serious deficiency in helping us understand the universality of sloppiness because we can get no statistics. We can not generalize from the specific properties of the Hilbert or Identity matrices but if we had an ensemble of sloppy models we could investigate properties such as the relationship between mean level spacings and fluctuations, or the typical performance of a particular algorithm. For these reasons we will now explore ensembles of models where we can study a whole

distribution of behavior with respect to sensitivities.

1.6   Wishart Statistics

While a large number of models are sloppy, not all multiparameter models share this behavior. For example, suppose we take the elementary multiple linear regression model for a single measurement y that depends on N independent variables a_i weighted by parameters p_i:

$$y_{\text{lin}}(\mathbf{a}, \mathbf{p}) = \sum_{i=1}^{N} p_i a_i = \mathbf{p} \cdot \mathbf{a}. \tag{1.19}$$

If we have K data points $y^{(k)}$ for variables $\mathbf{a}^{(k)}$, our cost is thus

$$C_{\text{lin}}(\mathbf{p}) = \sum_{k=1}^{K} (\mathbf{p} \cdot \mathbf{a}^{(k)} - y^{(k)})^2. \tag{1.20}$$

Linear correlation models like this are in essence fitting a plane to a cloud of K points in an N-dimensional space. The Hessian is

$$H_{ij} = 2 \sum_{k=1}^{K} a_i^{(k)} a_j^{(k)} \tag{1.21}$$

which is, up to normalization by the number of data points and subtracting off mean values (for this model the means are 0 because the cloud is centered at the origin), the sample covariance matrix of the data, $H = 2A^\top A$ where A is the $K \times N$ matrix of data points. A vital component of this model is that the N parameters are truly uncorrelated. The formalism can be generalized to include correlations between parameters but the standard analyses assume that the true covariance matrix is the Identity matrix. This class of matrices is known as the Wishart ensemble in the statistics community [36] and the Laguerre ensemble in the random matrix theory community [3]. For fixed $c = N/K$, the Wishart density
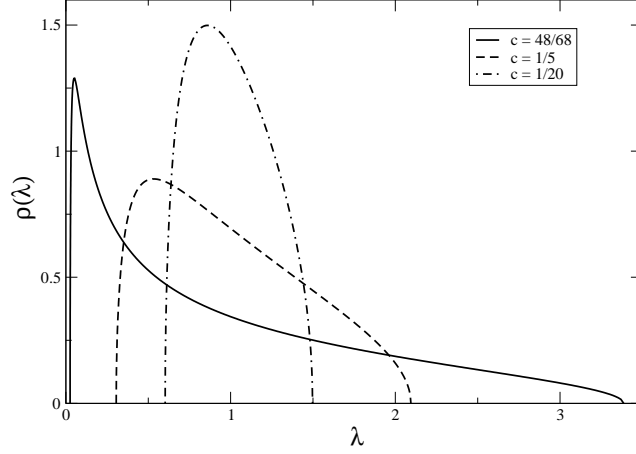
Figure 1.8: The Marčenko-Pastur distribution. The distribution is plotted for three values of c, the ratio of dimensions (parameters) to data points. These are the probability densities for the eigenvalues of Wishart matrices with the Identity matrix as the true covariance matrix. The value of $c = 48/68$ is chosen to mimic the number of parameters and data points for the PC12 model described in Figure 1.4 (a). The ratio b_+/b_-, the total range of eigenvalues, is 133, 6.9, and 2.5 for $c = 48/68$, 1/5, and 1/20 respectively.

of eigenvalues in the limit $N \to \infty$ is bounded between $b_\pm = (1 \pm \sqrt{c})^2$ and is known as the Marčenko-Pastur distribution [21]:

$$\rho(\lambda) = \max(0, 1 - \frac{1}{c}) \, \delta(\lambda) + \frac{\sqrt{(\lambda - b_-)(b_+ - \lambda)}}{2\pi\lambda c} I_{[b_-, b_+]} \qquad (1.22)$$

where I is the indicator function (zero outside the specified range and one within). For $N > K$ the linear correlation model has $N - K$ strictly zero singular values for the trivial reason that the system is underdetermined.

An example of the eigenvalues of a Wishart matrix is shown in Figure 1.4 (h) for the same numbers of parameters and data points as are in the PC12 model of column (a). The Marčenko-Pastur distribution for three values of c is shown
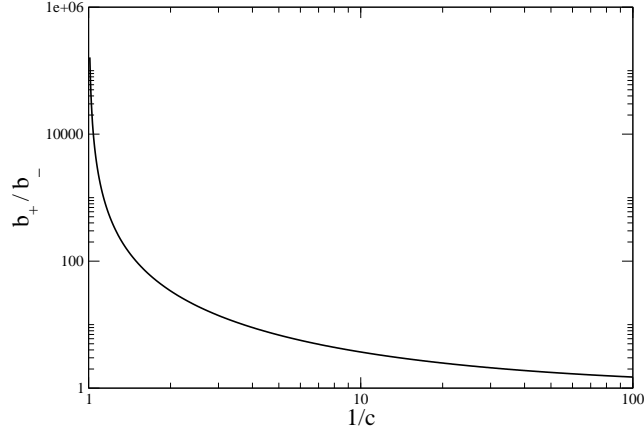
Figure 1.9: Total eigenvalue range for Marčenko-Pastur distribution. The overall range of eigenvalues, b_+/b_-, as a function of $1/c = K/N$, the amount of data relative to the number of parameters. When the number of data points is very close to the number of parameters, the Wishart ensemble is ill-conditioned but it is not truly sloppy because the range of eigenvalues is very sensitive to the amount of data.

in Figure 1.8, where it is clear that the distribution of eigenvalues becomes much tighter as the ratio of data points to parameters increases. Figure 1.9 depicts explicitly how the overall range of eigenvalues scales with the ratio of data points to parameters—the Wishart distribution is ill-conditioned only when the number of data points approaches the number of parameters. The overall range of eigenvalues in our sloppy models remains large even as data become available for all species at all times [5]; while the entire spectrum shifts upward with increasing amounts of data (both eigenparameters and bare parameters are more tightly constrained than they were before), the spectrum never flattens out (the asymmetry between stiff and sloppy directions in parameter space never disappears).

The Wishart family of distributions has two parameters - c, the ratio of data

points to dimensions, and V, the 'true' covariance matrix for the system. Traditional analyses of the Wishart matrices, such as that leading to the Marčenko-Pastur distribution, assume that $V = I$, the Identity matrix. This is the reason that as more data is collected and c decreases, the distribution of eigenvalues becomes sharply peaked around 1. Using the Identity matrix as the covariance matrix is equivalent to assuming that the parameters are truly uncorrelated and equally significant for the behavior of the model. When there are only a few more data points than parameters, randomness will cause the cloud of points to be slightly more extended in some directions than in others, and the covariance of the sampled points will be somewhat ill-conditioned. This is reflected in the relatively broad shape of the Marčenko-Pastur distribution for values of c near one. As more data points are collected however, the fact that there is no inherent broken symmetry in the system means that the cloud quickly resolves itself into a sphere, each direction in parameter space is equally well determined, and the eigenvalue distribution tightens quickly about one. This has little relevance for sloppy systems however because the parameters are neither uncorrelated nor equally significant for the behavior of the model.

It should be noted that the Wishart ensemble may still be useful in studying sloppiness. It would be instructive to carry out an analysis of the Wishart distribution with a sloppy true covariance matrix as a function of c. Anecdotally we observe that not very much data is needed to obtain a reliable picture of the eigenvalue spectrum, that especially the stiffest eigenvectors and eigenvalues become very well determined with modest amounts of data. This is consistent with the classic Wishart result shown in Figure 1.9 that the spectrum does not change significantly after a decent amount of data is available.

1.7 Random Matrix Theory

We were inspired to look for universality among sloppy models by the successes of random matrix theory (RMT), where similarities in eigenvalue plots like those in Figure 1.4 motivate the mathematical analysis of a well-defined ensemble of random matrices describing systems in disparate fields [22, 31].

The only immediately obvious properties of sloppy Hessians are that they are symmetric and that they have real elements. If this were all that was needed to define the ensemble of sloppy models, the sloppy spectra would mirror that of the Gaussian Orthogonal Ensemble (GOE) [22]. Figure 1.4 (e) shows the eigenvalues of a 48×48 member of the GOE; the eigenvalues are confined to a total range of two decades and are clearly not sloppy. The other two standard ensembles of RMT, the Gaussian Unitary Ensemble (GUE) and the Gaussian Symplectic Ensemble (GSE), also fail to explain the hallmark of sloppiness: exponentially large ranges of eigenvalues. Products of random matrices [3] (such as those describing electron transport through disordered wires) do have universality classes with singular values that are distributed roughly evenly over many decades (with more decades for longer wires) as shown in Figure 1.4 (f). While this quality is shared with sloppy models, we shall see later that definitive statistics of level spacings for the products of random matrices ensemble are not seen in sloppy models.

One of the most exciting results of RMT is the phenomenon of level repulsion: neighboring eigenvalues 'repel' one another such that the probability of a given spacing between levels vanishes as the spacing goes to zero. For the GOE, GUE, and GSE, this probability goes to zero as a linear, quadratic and quartic function of the spacing, respectively [31]. For the products of GOE matrices, a related result is that the variance of the spacing distribution is proportional to the mean

spacing [3]. As seen in Figure 1.10, level repulsion in sloppy models is qualitatively different from each of these predictions.

When the true parameters are widely spaced there exists no level repulsion; as these parameters become more and more similar the probability of small spacings between neighboring eigenvalues vanishes completely. The sharply peaked distribution for $\epsilon = 0.3$ is reminiscent of the rigid spectrum of Figure 1.4 (c) while the nearly Poisson statistics of the $\epsilon = 30$ case reflect the random spacings visible in Figure 1.4 (d). Indeed, this complete lack of level repulsion is comparable to what one would find with a superposition of several uncoupled sloppy models each with fewer exponentials drawn from a smaller original distribution, as if well-separated decay rates belonged to independent experiments. Once we derive a bound for the eigenvalues we will see that the sloppy model ensemble actually can have indefinitely strong level repulsion: the distribution of spacings between neighboring eigenvalues becomes a delta function as the average spacing grows and the system becomes sloppier.

We now know that clearly sloppiness is not a phenomenon described the classic ensembles of RMT. The approach of searching for an ensemble of matrices which have the correct statistical properties of sloppy systems is still valid, we simply need to discover for ourselves what defines this ensemble.

1.8   Vandermonde Ensemble

To form strong conclusions about sloppy models we must establish criteria sufficient to exclude the large variety of multiparameter systems that will not be sloppy. First, we specialize to models where the cost is a sum of squared residuals $C(\mathbf{p}) =$
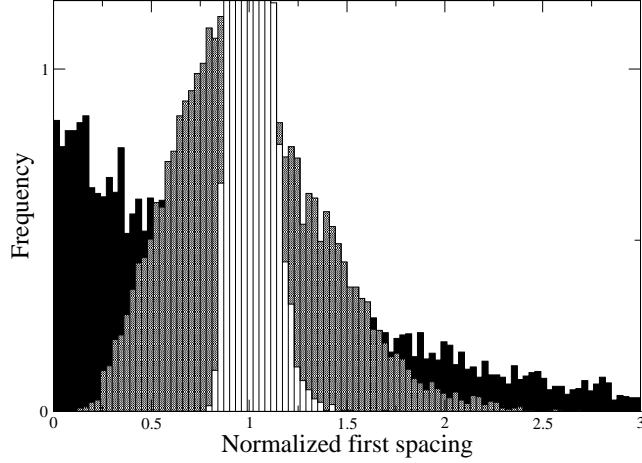
Figure 1.10: First eigenvalue spacing for fitting exponentials. The relative spacing between the stiffest and second-stiffest eigenvalues $(\log(\lambda_1/\lambda_2))$ for three versions of the fitting exponentials model, normalized to have both an integral and a mean of one. Each model is a sum of ten exponentials with all initial amounts fixed at one and 'data' distributed evenly in logarithmic time. In each instance the 'true' decay rates were generated from a log uniform distribution centered around 0 with width 2ϵ $(\ln(\gamma) = \mathrm{U}(-\epsilon, \epsilon))$. Ensembles of size 10,000 were generated with $\epsilon = 30.$, 3., and 0.3 for the widest, middle and narrowest distributions, respectively. Note that for the narrowest distribution the level repulsion has become so strong that there is virtually no probability of a level spacing less than 75% of the mean. Conversely, the distribution for $\epsilon = 30$ exhibits no significant level repulsion and in fact is almost Poisson.

$\sum_m r_m^2(\mathbf{p})$, where the sum may be continuous (e.g., an integral over time) and $r_m(\mathbf{p}) = y_m(\mathbf{p}) - d_m$ is the deviation of theory $y(\mathbf{p})$ from the experimental datum d_m. All of the empirical evidence we have for sloppiness so far comes from cost functions of this type. Many optimization problems that do not share this structure still maximize/minimize a scalar which reflects the conjoining of many competing, necessary factors and we would expect such problems to be sloppy as well. The sum-of-squares requirement also translates into a very concise structure for the Hessian which is precisely the object we need to study.

Second, to avoid including systems where each parameter is the subject of a separate experiment isolating that component, we make the (strong) assumption that all of the residuals $r_m(\mathbf{p})$ depend on the parameters $\mathbf{p}$ in a symmetric fashion (i.e., permuting $\mathbf{p}$ leaves r_m unchanged). Thus the residuals can be written in terms of symmetric polynomials of the parameters. The Newton-Girard formulas then provide a transformation to recast the residuals into the basis of power sum polynomials of the parameters, $r_m(\tilde{\mu}_1, \tilde{\mu}_2, \dots), \tilde{\mu}_k = \sum_{i=1}^{N} p_i^k$, which can also be viewed as the moments of the parameter distribution. Permutation symmetry is obeyed by our fitting exponentials problem but is violated by polynomial fits and the real world systems. We have seen, however, that the different polynomials have almost equivalent shapes in fitting the data and that this similarity is likely the source of sloppiness. In the biological and variational wave function examples, many of the basis functions are also quite similar in functional form.

Third, we found in fitting exponentials that the overall range of the bare parameters played an important role in sloppiness, smaller ranges lead to sloppier systems. To have control over this aspect of the model we will assume that the parameters are all confined to a small range $p_i \in [\bar{p} \pm \epsilon]$. Thus we define $\epsilon_i = p_i - \bar{p}$.

If we now consider $\tilde{\mu}_k = \sum_{i=1}^N (\bar{p} + \epsilon_i)^k$ we see that we can expand each term of the sum as $(\bar{p} + \epsilon_i)^k = \sum_{l=0}^k \binom{k}{l} \bar{p}^{k-l} \epsilon_i^l$ and factor out the contribution from $\bar{p}$ to the residuals. The kth parameter power sum can then be rewritten in terms of the $k + 1$ ϵ power sums of equal or lesser degree

$$\tilde{\mu}_k = \sum_{l=0}^k \binom{k}{l} \bar{p}^{k-l} \mu_l \tag{1.23}$$

where $\mu_l = \sum_{i=1}^N \epsilon_i^l$ and we are left with the residuals as functions of strictly the moments of the distribution of ϵ.

Conclusion 1 *For a cost function which is a sum of squared residuals $C(\mathbf{p}) = \sum_m r_m^2(\mathbf{p})$, if each residual $r_m(\mathbf{p})$ is a symmetric function of the parameters $p_1, ..., p_N$ and if the parameters are confined to a range $p_i^{(0)} \in [\bar{p} \pm \epsilon]$, then the Hessian matrix $H_{ij} = \partial^2 C/\partial p_i \partial p_j|_{\mathbf{p}^{(0)}}$ can be decomposed into*

$$H = V^\top A^\top A V \tag{1.24}$$

where the elements of A are bounded as $\epsilon \to 0$ and V is the Vandermonde matrix, $V_{kj} = \epsilon_j^{k-1}$.

In general the Hessian is

$$H_{ij} = \sum_m \left(\frac{\partial r_m}{\partial p_i} \frac{\partial r_m}{\partial p_j} + r_m \frac{\partial^2 r_m}{\partial p_i \partial p_j} \right) \tag{1.25}$$

but for the correct model at the true parameters the cost is zero, so $r_m = 0 \ \forall m$ and $H = J^\top J$ with the Jacobian

$$J_{mj} = \frac{\partial r_m}{\partial p_j} = \sum_{k=1}^K \frac{\partial r_m}{\partial \mu_k} k \epsilon_j^{k-1} = A_{mk} V_{kj} \tag{1.26}$$

$$A_{mk} = \frac{\partial r_m}{\partial \mu_k} k \tag{1.27}$$

$$V_{kj} = \epsilon_j^{k-1} \tag{1.28}$$

where K is the maximum degree (possibly ∞) to which we expand in ϵ. Thus $H = J^\top J = V^\top A^\top A V$. □

Here V, the famous Vandermonde matrix, is the heart of the sloppy model universality class. Reminiscent of random matrix theory ensembles, we are now interested in the Vandermonde ensemble of Hessians of the form $V^\top A^\top A V$. The Vandermonde matrix is well-known primarily because its determinant (for $N = K$) can be expressed analytically, $\det(V) = \prod_{i<j}(\epsilon_i - \epsilon_j)$. As $\epsilon \to 0$ this product is tiny, $\det(V) = \mathcal{O}(\epsilon^{N(N-1)/2})$. While the elements of A do, in general, depend on the parameter values, they either approach a constant or zero in this limit, so $\det(A)$ remains finite as $\epsilon \to 0$. Hence, the determinant of H, $\det(H) = \det(V)^2 \det(A)^2$ is also tiny as $\epsilon \to 0$, so the product of the eigenvalues of H is small. As we saw with the Hilbert matrix and fitting monomials to data (Section 1.5), the Hessian can be viewed as the square of the transformation between the bare parameters and the eigenparameters (equation 1.17), and transformation matrices with very small determinants are a signature of sloppy models.

To show that the eigenvalues in our Vandermonde ensemble are evenly spread in logarithm, we will make use of an apparent truth about matrices:

Conjecture 1 *Let $S \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Let $E \in \mathbb{R}^{n \times n}$ be diagonal with $E_{ii} = \epsilon^{i-1}$ and $0 < \epsilon \leq 1$. Then the mth largest eigenvalue of ESE is bounded by $\lambda_m = \mathcal{O}(\epsilon^{2(m-1)})$.*

We have two reasons to believe this conjecture is true. (1) This is a self consistent combination of proof by induction and perturbation theory. Let matrices S and

E be $n \times n$ and assume the eigenvalues of ESE scale as $\lambda_m = l_m \epsilon^{2(m-1)} + \mathcal{O}(\epsilon^{2m})$ for some nonzero coefficient l_m that does not depend on ϵ. Now consider adding a row and a column to E and S. Treating this addition as a perturbation on the old system, the corrections to λ_i, one of the previous n eigenvalues, scale as $\epsilon^{2(n+1)+2i-4}/(\epsilon^{2i-2} - \epsilon^{2(n+1)-2}) \approx \epsilon^{2n}$ which is a small perturbation. The new eigenvalue, λ_{n+1}, is also given by perturbation theory and it is of order $\epsilon^{2(n+1)-2}$. This is precisely the scaling form we had for the previous n eigenvalues, so starting induction from the fact that a 1×1 system has eigenvalue $S_{11}\epsilon^0$, we see that the proof is self-consistent. (2) Extensive numerical tests (Appendix A) show an even sharper result: the mth largest eigenvalue, λ_m, is bounded above by the mth largest row sum of EES, where the row sum for row k is $r_k = \sum_l \epsilon^{2(k-1)}|S_{kl}|$. Since $\|S\|_\infty = \max_k(\sum_l |S_{kl}|)$ this implies that $\lambda_m \leq \|S\|_\infty \epsilon^{2(m-1)}$ and switching to big O notation, $\lambda_m = \mathcal{O}(\epsilon^{2(m-1)})$. $\qquad\square$

This conjecture implies a remarkable apparent fact about the eigenvalues of any symmetric, positive definite matrix. That this has not been discovered before is quite surprising since the eigenvalues of symmetric, positive definite matrices are of great importance in many fields and efficient ways of bounding their values are of significant use.

Corollary 1 *Let $S \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Let the sum of the absolute value of the elements in the ith row be called the ith row sum, $r_i = \sum_j |S_{kl}|$. Then the sorted eigenvalues of S are each bounded above by their corresponding row sum, $\lambda_i \leq r_i$.*

This follows from Conjecture 1 with $\epsilon = 1$. $\qquad\square$

We have numerical evidence that to leading order in ϵ the eigenvectors of the Hessian are the right singular vectors of the Vandermonde matrix (Appendix B).

(A non-square matrix V has a singular value decomposition $V = U\Sigma W^\top$, where the non-square matrix Σ has the singular values of V along the diagonal and is zero elsewhere; W gives the right singular vectors, which are eigenvectors of $V^\top V$, and U gives the left singular vectors, eigenvectors of $VV^\top$.) Motivated by this numerical evidence, we shall transform our Hessian to this basis. This is an interesting result because information about the model is encoded in the matrix A, not the Vandermonde matrix (or its eigenvectors). The fact that the eigenvectors of the Hessian are not strongly determined by the elements of A means that attempting to interpret the composition of the eigenvectors (particularly the sloppy ones) may not provide much insight into the model itself. Our experience with sloppy models is precisely this, that the exact composition of the eigenvectors is not usually not very revelatory.

We first bound the singular values of the Vandermonde matrix. Conveniently, $VV^\top$ has the form necessary for Conjecture 1.

Conclusion 2 *The mth-largest singular value $\sigma_m(V)$ of the N-column Vandermonde matrix, $V_{ij} = \epsilon_j^{i-1}$ is $\mathcal{O}(\epsilon^{m-1})$.*

The singular values of V are the positive square root of the eigenvalues of $VV^\top$. Factoring the appropriate power of ϵ from each row of the Vandermonde matrix gives $V = EX$ and $VV^\top = EXX^\top E$ where E is the same as in Conjecture 1 and the elements of X are bounded by one. Equating $XX^\top$ with the matrix S in Conjecture 1, we conclude that the eigenvalues of $VV^\top$ scale as $\lambda_m(VV^\top) = \mathcal{O}(\epsilon^{2(m-1)})$ and thus $\sigma_m(V) = \mathcal{O}(\epsilon^{m-1})$. $\qquad\square$

We now transform the Hessian into this basis, and again use Conjecture 1 to bound its eigenvalues.

Conclusion 3 *The eigenvalues of the Hessian matrix for the class of models in Conclusion 1 scale as*

$$\lambda_i(H) = \mathcal{O}(\epsilon^{2(i-1)}) \tag{1.29}$$

Starting with the decomposition $H = V^\top A^\top A V$, taking the singular value decomposition of $V = U\Sigma W^\top$, and transforming the Hessian into the basis of the right singular vectors of the V, we have $W^\top H W = \tilde{H} = \Sigma^\top U^\top A^\top A U \Sigma$. By Conclusion 2 we know that $\Sigma_{ii} = \mathcal{O}(\epsilon^{i-1})$. By construction the elements of A are well-behaved as $\epsilon \to 0$ and since U is an orthogonal matrix its elements too cannot diverge in this limit. This means that $\tilde{H}_{ij} = \mathcal{O}(\epsilon^{i+j-2})$. By Conjecture 1 we know that $\lambda_i(\tilde{H}) = \mathcal{O}(\epsilon^{2(i-1)})$ and since $\tilde{H}$ is simply an orthogonal transformation of H, $\lambda_i(H) = \mathcal{O}(\epsilon^{2(i-1)})$. $\qquad\square$

While rigorous universality is only expected as the system size approaches infinity, we find empirically that models with more than roughly ten parameters are often recognizably sloppy.

Since all polynomial fits in the basis of monomials have the Hilbert matrix as the Hessian, even very small systems have a wide range of eigenvalues (the eigenvalues of the 3×3 Hilbert matrix are $\lambda \approx \{1.4, 0.12, 0.0027\}$. In fitting exponentials, a small system can be quite sloppy provided the true parameters are from a sufficiently narrow range. In an analysis of density-functional theory (DFT) calculations with only three parameters, the eigenvalues spanned orders of magnitude: $\lambda \approx \{1876, 44, 0.5\}$ [23, 18]. While there are clearly two orders of magnitude between consecutive eigenvalues for this model, we would hesitate to claim that all three parameter models will be sloppy. In a wide selection of biological network models with numbers of parameters ranging from eight to one hundred forty three, each model had a characteristically sloppy sensitivity spectrum [16]. We can

further investigate the relationship by taking relatively large models and simply restricting the number of free parameters(Appendix C) and again, models with as few as eight parameters are recognizably sloppy.

Do these results tell us anything about the statistics of level spacings? Unless two parameters are strictly equal or the residuals are independent of a particular moment of the parameter distribution, Conclusion 3 shows that $\lambda_i = l_i \epsilon^{2(i-1)} + \mathcal{O}(\epsilon^{2i})$ for some non-zero coefficient l_i. The relative spacing between neighboring eigenvalues, to first non zero order, is $s_i = \log(\lambda_i/\lambda_{i+1}) = \log(l_i/l_{i+1}) - 2\log\epsilon$. Figure 1.11 depicts the accuracy of this relationship. For a fixed model but an ensemble of random parameters, the distribution of coefficients l_i has a finite width as $\epsilon \to 0$. Therefore the distribution of s_i over the ensemble, normalized by $2\log\epsilon$ such that the average spacing is unity, goes to one with a width which vanishes as $\epsilon \to 0$, as is illustrated in Figure 1.10. This means that the whole system is becoming not only more sloppy (larger spacing) but it is becoming almost deterministically so (strong level repulsion). Figure 1.4 (c) is a manifestation of this rigid spacing between levels due to remarkably strong level repulsion. It should be noted that the calculations necessary to confirm these predictions (reliably calculating remarkably small eigenvalues) were only possible because of the ability to set arbitrarily high precision and accuracy in Mathematica.

1.9   Vandermonde Decompositions

What is the link between the Vandermonde ensemble at small ϵ and the behavior of real world sloppy models (Figure 1.4 columns (a), (b)) and the behavior at large ϵ (column (d))? These latter systems share the roughly uniform density of log-
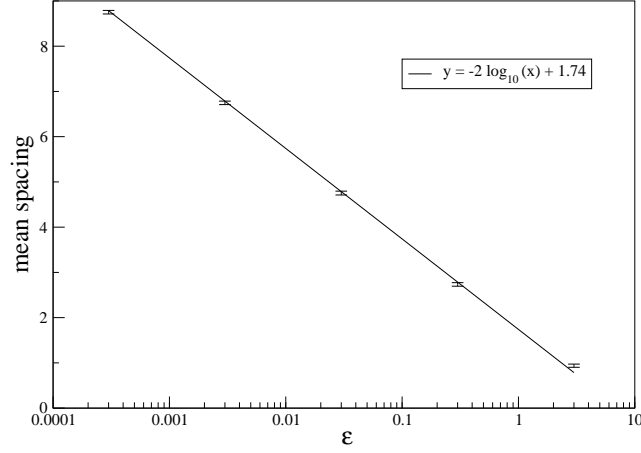
Figure 1.11: Mean log level spacing. Plotted is $\log_{10}(\lambda_i/\lambda_{i+1})$, averaged over the all spacings, as a function of ϵ. The model is a sum of ten exponentials with all initial amounts fixed at one. In each instance the 'true' decay rates were generated from a log uniform distribution centered around 1 with width 2ϵ. Ensembles of size 1500 were generated and the mean spacing (averaged over the ensemble and over all nine spacings) is plotted with standard errors. Note the excellent fit of $y = -2\log_{10}(x) + b$ where only b is allowed to vary; $\lambda_n \propto \epsilon^{2n}$.

eigenvalues over many decades that is the signature of sloppy models but do not exhibit strong level repulsion. The real world models also do not share the strict requirement that the residuals be perfectly symmetric functions of the parameters. We conjecture that while not all of the parameters are interchangeable in real world sloppy models, there are Vandermonde subsystems lurking below the surface. Thus the fastest decay rates in Figure 1.4 (d) constitute one Vandermonde subsystem and the slowest decay rates another. Indeed, the Poisson statistics of level spacings when fitting exponential decays from a wide range can be reproduced by superimposing the spectra of several separate experiments, each fitting decays from a narrower range (e.g. the level spacing statistics for fitting forty nine exponentials with decay rates in the wide range $2\epsilon = 50$ as in Figure 1.4 (d) are equivalent to the level spacing statistics when seven separate experiments are superimposed over one another, each fitting seven exponentials with decay rates in the narrow range $2\epsilon = 3.5$ as in Figure 1.4 (c)). Such a decomposition into Vandermonde subsystems is also illustrated by modifying the net radiation model to include the initial amounts of the elements as unknown parameters (Section 1.4). Now the parameters clearly separate into two classes – decay rates and initial amounts. Each class alone fits the assumptions of the Vandermonde ensemble, produces rigidly (strong level repulsion) sloppy spectra, and generates nearly equivalent patterns of changes in the residuals. When mixed together however, the fact that parameters from one class can not compensate for parameters of the other class destroys the correlations between levels and they do not repel each other anymore as is evident in Figure 1.12. Similarly, a full many body wave function in quantum Monte Carlo [24] is composed of the sloppy space of the Jastrow parameters in figure 1.4 (b) and a non-sloppy subspace of the Configuration Interaction coefficients

describing single-particle orbitals.

These results motivate algorithms for the decomposition of real world sloppy models into rigidly sloppy Vandermonde subspaces whose components are effectively redundant. Such a decomposition would be useful for three separate reasons: a) explaining why a particular model is sloppy overall, b) suggesting routes for model reduction and coarse graining by subsuming degrees of freedom within Vandermonde systems, and c) prescribing changes in parameters to alter specific aspects of model behavior.

It is instructive to consider the structure of a composition of distinct Vandermonde subsystems. Let H_A be the Hessian for one Vandermonde system, H_B be the Hessian for a second subsystem and let the matrix M define the coupling between them. Without loss of generality we can assume both Hessians are diagonal but we can not assume anything about the structure of M. The Hessian of the coupled system, H_{A+B} is then

$$H_{A+B} = \begin{pmatrix} H_A & M \\ M^\top & H_B \end{pmatrix} \tag{1.30}$$

$$= \left(\begin{array}{ccc|ccc} \ddots & & 0 & & & \\ & \lambda_A & & & M & \\ 0 & & \ddots & & & \\ \hline & & & \ddots & & 0 \\ & M^\top & & & \lambda_B & \\ & & & 0 & & \ddots \end{array} \right). \tag{1.31}$$

Clearly the structure of M is significant for the sloppiness of the composite system. Surely some types of coupling could destroy the sloppiness altogether. While future analysis may be rewarded by tackling the problem of coupling Vandermonde subsystems from this angle, we will take a different approach and instead attempt
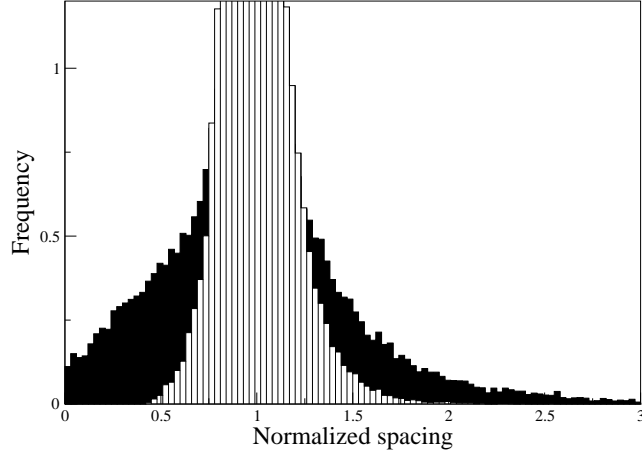
Figure 1.12: Fitting exponentials with and without level repulsion. The relative spacing between the eigenvalues ($\log(\lambda_i/\lambda_{i+1})$, normalized to the mean value for each spacing) for two versions of the fitting exponentials model, normalized to have an integral of one. Each model has nine parameters (each selected from $\ln(p_i) = \mathrm{U}(-0.3, 0.3)$) — in the case of the wide distribution there are four initial amounts (plus one fixed initial amount) and five decays rates but for the narrow distribution there are ten fixed decay rates, one fixed initial amount and nine initial amounts that can vary. The distribution for fixed initial amounts but unknown decay rates is very similar to the narrow distribution here (see Figure 1.10) and is therefore not included. Note the much stronger level repulsion when all parameters are of the same type (just initial amounts) than when there are distinct subclasses of parameters (initial amounts and decay rates).

to study how sloppy subspaces arise in specific, known cases.

We have begun the search for such an algorithm but it is clearly a deep enough problem to warrant much research beyond this thesis. The key suggestion from the Vandermonde ensemble is that evenly spaced log eigenvalues should be intimately linked to parameter redundancy for specific aspects of model behavior. Attempting to pick low hanging fruit, we take the problem of fitting both initial amounts and decay rates in a multi-exponential model of radioactive decay. We know that each problem separately affects the residuals in a different way and that each problem separately has rigidly sloppy spectra. When we attempt to fit both parameters simultaneously, however, the eigenvalues lose all their repulsive qualities and the statistics of level spacings is nearly Poisson. We would like a function to optimize which quantifies either how 'rigidly sloppy' a given Hessian is or how redundant a set of parameters are on the residuals (or a combination of both).

We will first tackle the problem of finding subspaces with rigidly sloppy spectra. Given a particular Hessian, we would like to reorient our basis such that the subspace Hessians determined by some black diagonal structure are each their own Vandermonde style system. In general we will not know how many subspaces we are looking for or what the dimension of each subspace is, but for our expanded fitting exponentials model we know that there are two Vandermonde subspaces and that they have equal dimensions (there are the same numbers of decay rates and initial amounts). For an $N \times N$ Hessian, we are therefore searching for the $N \times N$ orthogonal transformation which maximizes the sloppiness of the two $N/2 \times N/2$ submatrices along the diagonal. There is no obvious single formula to measure the sloppiness of a given spectra. We want something that favors both a large overall range of eigenvalues and also identical spacings between each level. Just

how to balance these different qualities is not clear, and given the array of possibilities the best approach is to simply implement each criteria and judge how well it works. For a matrix H with eigenvalues $\lambda(H)$ we have found that the sum of ratios of all neighboring eigenvalues reliably produces the sloppiest spectra, $S(H) = \sum_{i=1}^{N-1} \lambda_{i+1}/\lambda_i$. The difference between this criteria and simply the total range between the stiffest and sloppiest eigenvalues is a subtle one but remarkably effective. The desire for equal spacings is lost when simply optimizing the overall range, the optimized spectra do indeed have very well separated largest and smallest eigenvalues but the levels in between are scattered randomly. Optimizing instead the sum of all relative spacings produces very regular spectra. Since we are looking for a function to minimize, we choose the ratios with the larger eigenvalue in the denominator as opposed to the inverse.

The next question is whether this also identifies sets of parameters which determine the same features of model behavior. Again we are confronted with the problem of how to quantify this quality and again we resort to empirical studies for the answer. One certainty is that this is a question about the Jacobian matrix and not the Hessian because we are looking for particular patterns in the residuals which have been averaged out by the Hessian. Recall that each element of a Jacobian corresponds to the derivative of a residual with respect to a parameter—$J_{ij} = \partial r_i/\partial p_j$. Summing the dot products of neighboring columns of the Jacobian is the measure which functions best in our empirical tests. A dot product of zero between two columns means that changes in those reoriented parameters have completely distinct patterns of effects on the residuals. A dot product of one (each row is normalized to have length one to avoid interference from trivial rescaling of rows) means that the two directions in parameter space produce identical effects

on the residuals. We thus use $R(J) = \sum_{j=1}^{N-1} \left(1 / \sum_{i=1}^{N} J_{ij} J_{ij+1}\right)$ as our measure of how redundant successive parameter directions are.

In each case we are optimizing over the space of orthogonal matrices. There are several ways to continuously parameterize the space of orthogonal matrices. One could take the matrix exponential of all anti-symmetric matrices or use a procedure such as SVD or Gram-Schmidt to find orthogonal basis of general matrices but we find that the Cayley transformation of anti-symmetric matrices performs the best. It is quicker to compute than the matrix exponential of an anti-symmetric matrix and it has half the parameters of finding an orthogonal basis (e.g. by Gram-Schmidt or SVD) for general matrices. The Cayley transformation takes an anti-symmetric matrix, $A = -A^{\top}$, and produces an orthogonal matrix U by $U = (I + A)(I - A)^{-1}$ where I is the Identity matrix.

Figure 1.13 demonstrates that of all the possible decompositions, separating the decay rates from the initial amounts produces the sloppiest subspaces. In this example, H is the Hessian matrix organized to have the block structure of Equation 1.9. The matrix $P^{\top} H P$ is a permutation of H where the block structure has been disrupted by swapping the row and column for one of the decay rates with a row and column for one of the initial amounts. Starting from this permuted Hessian we then used the Cayley transformation to minimize $S(U_1^{\top} P^{\top} H P U_1)$ to find the optimal orthogonal matrix U_1. We also did a second optimization starting from H to find the orthogonal matrix U_2 which minimizes $S(U_2^{\top} H U_2)$. This procedure was performed one thousand times, each time beginning with a random set of decay rates and initial amounts. The mean and standard deviations over this ensemble of one thousand instances are depicted in the figure. If the optimization routines always found a single, best optimum then the values for $S(U_1^{\top} P^{\top} H P U_1)$
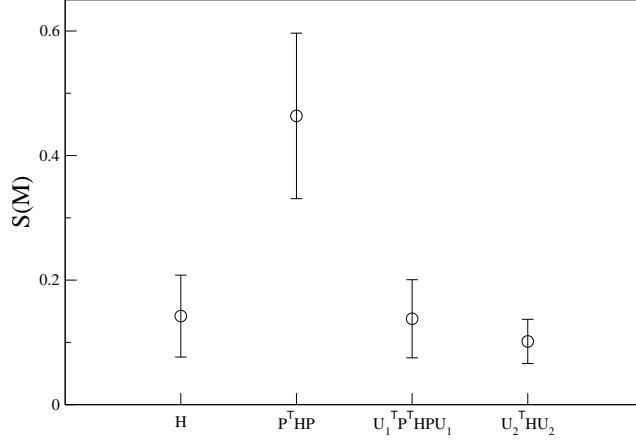
Figure 1.13: Sloppiness of subspaces in fitting exponentials. The sloppiness of various subspaces of the fitting exponentials problem with both decay rates and initial amounts as measured by $S(M)$ for the matrices M listed on the horizontal axes. For an ensemble of one thousand random sets of initial amounts and decay rates, H is the Hessian matrix of form Equation 1.9, P is a permutation matrix which swaps the entries corresponding to one of the decay rates with those corresponding to one of the initial amounts, U_1 is an orthogonal matrix found by optimizing $S(U_1^\top P^\top H P U_1)$ and U_2 is an orthogonal matrix found by optimizing $S(U_2^\top H U_2)$.

and $S(U_2^\top H U_2)$ would be the same. The similarity of the two measures is reassuring that there are not significant local minima leading to incomplete convergence of the optimization algorithm. The fact that $S(P^\top H P)$ is substantially larger than $S(H)$ means that interchanging decay rates with initial amounts makes the two subspaces less like members of the Vandermonde ensemble. The fact that $S(H)$ is so similar to either of the optimized versions, $S(U_1^\top P^\top H P U_1)$ and $S(U_2^\top H U_2)$, means that even the best grouping possible is not substantially better than the naïve grouping.

In Figure 1.14 we see that optimizing the sloppiness of the subspaces, $S(H)$ does indeed correlate strongly with optimizing the redundancy of the parameters on specific aspects of model behavior, $R(J)$. The matrices P, U_1 and U_2 are from the optimizations described for Figure 1.13 but here they are applied to the Jacobian and the corresponding $R(M)$ for matrix M is calculated. We see that the decomposition into simply decay rates and initial amounts not only optimizes the sloppiness of the subspaces (Figure 1.13), it also optimizes the redundancy of the two sets of parameters since $R(J)$ is so small. The fact that the matrices U_1 and U_2, which were obtained solely by optimizing $S(M)$, also lead to noticeably reduced values of $R(M)$. The fact that $R(J)$ is consistently the lowest value underscores the usefulness of this test problem for testing such Vandermonde decomposition algorithms, since we reliably know what answer the algorithm should find.

The final difficulty is that blindly optimizing according to these criteria will in general produce results that are uninterpretable to a human being—the curse of dimensionality, the effects of noise, and general entropic reasons will lead to orthogonal transformations which do indeed optimize our objective function but which add little or nothing to our understanding of the system. We therefore prefer orthogonal matrices which are well-localized (contain just a few large elements) over more diffuse (many elements of roughly equal size) transformations. In order to bias our optimization procedure toward such well-localized matrices we add to the total cost function a measure called the *participation ratio* in quantum mechanics. Mathematically, the participation ratio, P, of a matrix M is given by $P(M) = \sum_i \left(\sum_j M_{ij}^4 / \sum_j M_{ij}^2 \right)$. In our case the denominator drops out because we are dealing with orthogonal matrices so the sum of squares of any row or column of the matrix is one. Skipping the derivation from quantum mechanics,
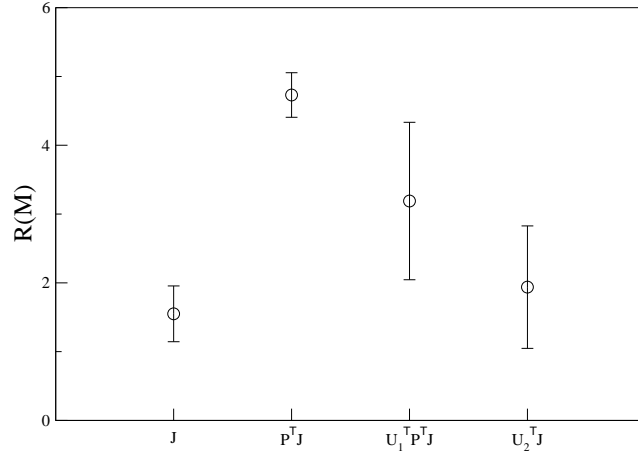
Figure 1.14: Measuring the redundancy of parameter subsets. The redundancy of various subsets of parameters in determining distinct aspects of model behavior as measured by $R(M)$ for the matrices M listed on the horizontal axes. The matrices P, U_1, and U_2 are from the optimization procedure described in Figure 1.13 while J is the Jacobian corresponding to the same sets of decay rates and initial amounts as the matrices H from that Figure. The ordering of decay rates and initial amounts in J is the same as for H.

simple inspection of this formula shows that it is a measure of the variance of the elements of M and hence is a minimum when all entries are of the same size and a maximum when there are a few big entries and many small entries. For this reason we want to minimize the inverse of the participation ratio.

Optimizing either the sloppiness of the subspaces ($S(H)$ or $R(J)$) or the participation ratio of the transformation matrix ($P(M)$) alone has proven to be quite simple and can be tackled with standard optimization algorithms. The problem of optimizing both criteria simultaneously (or optimizing one and then turning up the weight of the other measure) has proven very difficult however. The problem is that the participation ratio changes at a much slower rate as the parameters (entries of the anti-symmetric matrix in the Cayley transform) vary than does the sloppiness measure. This is demonstrated in Figure 1.15 which shows $S(H)$ and $P(U)$ as one parameter is changed. Note the wide discrepancy in scales, the variations in $S(H)$ are roughly three orders of magnitude larger than the scale of $P(U)$. A deeper understanding of these landscapes is necessary before any successful optimization can be accomplished. These hurdles are by no means unsurmountable, a clever algorithm should be able to optimize both the sloppiness and the localization.

1.10   Conclusion

Complex multiparameter models from a wide array of scientific fields are often *sloppy*: they each have an exponentially large range of sensitivities to changes in underlying parameter values. This occurs because the parameters natural for experimental manipulation or human description are often a severe distortion of the basis natural for describing system behavior.
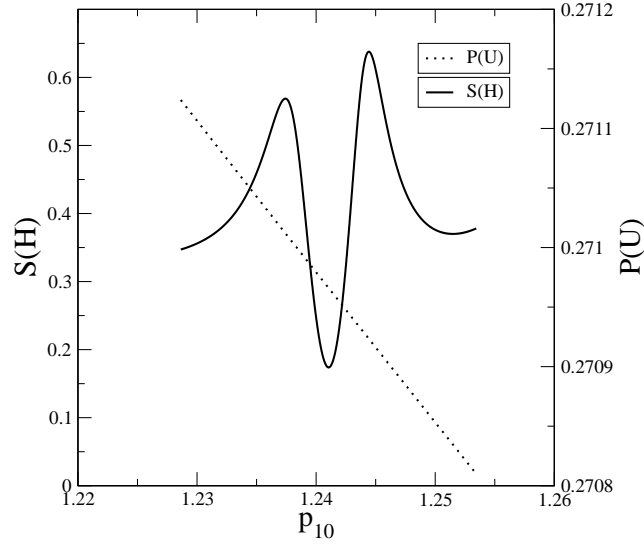
Figure 1.15: Correlating sloppy spectra with redundant parameters. The sloppiness of the resulting subspaces, $S(H)$, and the inverse participation ratio of the transformation matrix, $P(U)$, as one parameter (an element in the anti-symmetric matrix in the Cayley transform) is adjusted. The parameters were first optimized to minimize $S(H)$. The vastly different scales at which the two functions vary explains the difficulty of then turning on the participation ratio as a component of the objective function.

Sloppiness may at first seem to condemn complex multiparameter models as useless because it is so difficult to constrain the parameters. On the contrary, sloppiness is in fact a saving grace—only the small number of stiff combinations of parameters need to be constrained in order to restrict model behavior. The uncertainty in parameters along sloppy directions can be quite large and the model will still generate precise predictions. The large range of sensitivities necessitates that predictions with such complex models be accompanied by rigorous error bars due to any underlying parameter uncertainty. Experimental results lying outside such error bars are then strong evidence that something is structurally wrong with the model instead of simply inaccurate parameterization. For perfectly linear models in which the cost function is purely quadratic, parameter uncertainty can be propagated to prediction uncertainty through straightforward analytical calculations. The vast majority of complex models are not that simple however, and nonlinearities in the cost function quickly turn the sloppy directions into curved sloppy manifolds as shown in Figure 1.3. To account for such nonlinearities, a Markov Chain Monte Carlo (MCMC) approach is necessary to propagate errors [7]. In this approach an ensemble of parameter sets is generated which are each consistent with the current data. Predictions are then made simply by using each set of parameters in the ensemble and weighing their predictions by how well they fit the current data. The ability to gauge the stiff and sloppy directions in parameter space and to scale steps appropriately is vital for the success of the MCMC algorithm, as steps too large along stiff directions will never be accepted and steps too short along sloppy directions will never converge [7].

While sloppiness is not a statement about the quality of a model, it is a statement about how best to constrain model predictions. Since the eigenparameters

which naturally describe model behavior tend not to be aligned with the bare parameters, individually measuring the bare parameters is a very inefficient route to constraining model behavior. Collectively fitting all the parameters to previously measured system-wide data on the other hand naturally constrains parameters along stiff directions and allows large uncertainties only along sloppy directions [16]. While collectively fitting model behavior for a sloppy system will never allow one to reliably determine parameter values, the parameters of such models are often of little interest and are instead simply a means to the end of making precise predictions for future experiments.

The prevalence of sloppiness in complex models is of particular importance to the emerging field of (computational) systems biology. As the technology for making measurements such as sequencing genomes, measuring entire proteomic repertoires of cells, and imaging localization and transport, becomes more advanced and more high-throughput, researchers are becoming increasingly interested in the functioning of interconnected networks of biomolecules. The origins and universality of sloppiness are not tied to a particular mathematical framework (e.g. coupled ODEs) and are therefore relevant to any model that constitutes a convoluted mapping from parameters to system behavior. A worry in the field of computational biology is that highly parameterized models can never be constrained enough to offer useful insights [2]. The fact that such models are dominated by only a few stiff modes however means they can be quite valuable even early in the research program because nontrivial, falsifiable predictions can be made with a surprisingly modest amount of previous data and despite enormous parameter uncertainties.

The universality of sloppiness holds another exciting possibility for the field of biology. As opposed to models where the parameters are largely human constructs

to describe the system (e.g. fitting polynomials simply to find trends), models of the molecular interactions in complex biological networks are closely tied to the process of evolution. The random mutations in the DNA sequence through which much of evolution operates have direct consequences for the biochemical reactions in which the protein product participates. Similar consequences hold for regulatory DNA sequences or for sequences which code for functional RNA instead of protein. We have not explored this issue in any detail but sloppiness would seem to provide a novel structure for the fitness landscape through which evolution moves. The exponentially large range of sensitivities is one unique issue for evolution to tackle but the fact that eigenparameters tend not to be aligned with bare parameters is a more substantial, and interesting, problem. It is presumably much easier for evolution to take steps along the bare directions, but any given bare parameter usually includes at least some component along a stiff direction so any single step would be very costly. Of course we have no idea what 'cost function' any real organism is truly experiencing or has experienced. Instead, studying *in vitro* evolution, where the scientist determines what trait(s) to select for, could provide interesting insight to the role of stiff and sloppy directions in nature.

It should be noted that there is a large class of multiparameter optimization problems that have not been dealt with in this work. All the models considered have been sum-of-squares cost functions where many separate aspects of model behavior are being balanced. Many optimization tasks however have simply one figure-of-merit. We have not analyzed such systems empirically so we have no evidence whether or not they are sloppy. A sum-of-squares cost function is also one of the assumptions made to derive the Vandermonde ensemble, so those results simply do not apply to this second class of models. It does seem quite unlikely that

each parameter in such a system has an entirely unique effect on model behavior. Since parameter redundancy appears to be the heart of sloppiness, this argues that even single figure-of-merit models should be sloppy but much more work needs to be done before a definitive statement can be made.

Sloppiness is a general phenomenon whereby multidimensional nonlinear models exhibit exponentially large ranges of sensitivities. This affects the estimation of both statistical [4, 16] and systematic [30, 23] errors, favors collective parameter fitting over individual measurements [16], and motivates scale invariant algorithms for optimization. Understanding the origins and implications of sloppiness in its various incarnations offers new, fundamental insights into complex systems.

Chapter 2

Computational Model of Quorum Sensing in *Agrobacterium tumefaciens*

2.1   Introduction

Quorum sensing is the process by which bacterial cells regulate an important activity or property of the cell in response to changes in the population density. In general this process rests on the production of a signaling molecule (termed an 'autoinducer') whose extracellular concentration increases with the bacterial population density.

The range of behaviors regulated by quorum sensing in various bacteria is extraordinarily broad: bioluminescence in *Vibrio fischeri*, pathogenesis in *Staphylococcus aureus*, biofilm formation in *Pseudomonas aeruginosa*, and sporulation in *Bacillus subtilis* to name a few [25, 34]. In most bacteria gene expression is upreg-

ulated and not downregulated in response to a quorum but in some cases this is done by activation while in others it is through derepression.

Gram-positive bacteria that possess quorum sensing networks tend to use short peptides as the signaling molecules. These oligopeptides bind either histidine kinases in the cell membrane or are transported into the interior of the cell and bind a phosphatase, thereby activating a phosphorelay network. One interesting exception to this phosphorylation-based signaling occurs in *Enterococcus faecalis*. The peptide autoinducer in this bacterium is transported into the cell but then it activates transcription by directly binding a transcriptional repressor, relieving its activity [10].

In Gram-negative bacteria the autoinducers are predominantly *N*-acyl-homoserine lactones (AHLs) whose synthesis is catalyzed by a bacterial protein. This family of molecules is characterized by having a lactone ring joined to a carbon chain. The main variations between autoinducers are the length and saturation of the carbon chain and the oxidation state at the C-3 position. The majority of these molecules can freely diffuse through the membranes of cells although those with particularly long carbon chains rely on active transport by proteins in the membrane. At roughly micromolar concentrations, diffusion out of and into the cell is balanced and the autoinducers are detected by binding to transcription factors. In the majority of bacteria the autoinducer increases the activity of the transcription factor but in a few cases the autoinducer antagonizes the transcription factor.

In addition to these general rules for the molecular basis of quorum sensing (oligopeptide communication by Gram-positive bacteria and AHL based signaling in Gram-negative bacteria) there are of course several exceptions [25]. Due to the mechanism of a diffusible extracellular signaling molecule it is also possible for

'quorum sensing' networks to detect diffusion barriers as well as high population levels [29]. It is still possible to consider an abstract, simplified quorum sensing network that describes the majority of the known systems. Amongst the gene targets of the various quorum sensing networks, expression of the gene responsible for the autoinducer is often elevated, creating a positive feedback loop which drives the system to an 'activated' state. In the Gram-positive bacteria this would be the gene for the oligopeptide itself while in Gram-negative bacteria it would be the gene for the protein which catalyzes synthesis of the AHL. A simple picture of this idea for Gram-negative bacteria is depicted in Figure 2.1. The explanation of these simplified dynamics is depicted in Figure 2.2.

2.2 Quorum Sensing in *Agrobacterium tumefaciens*

One of the best characterized quorum sensing networks is that of the α- proteobacterium *Agrobacterium tumefaciens*. This bacterium typically lives in the area around the roots of trees and plants called the rhizosphere and is the causative agent of plant tumors called crown galls. *Agrobacterium* is primarily well-known because the basis of its virulence, the ability to transfer DNA into host cells, can be used as a valuable tool in the laboratory.

In *Agrobacterium tumefaciens*, quorum sensing begins with pathogenesis and pathogenesis begins when the bacteria (a) sense chemical signals from plant cells, especially from wounds in plants. These signals cause the bacteria to (b) initiate an infection. If the bacteria (c) sense that the infection is successful, they start the process of (d) counting a quorum. After the population becomes dense enough and
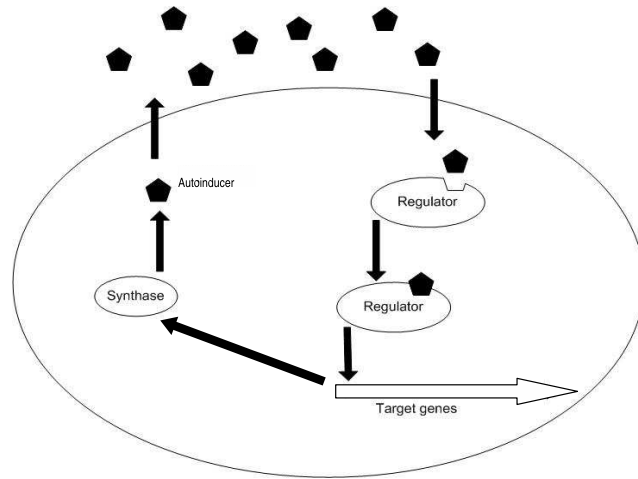
Figure 2.1: A simplified, generic quorum sensing network. The large oval represents a Gram-negative bacterial cell, smaller ovals represent proteins, pentagons represent small molecules, the outlined arrow represents bacterial DNA, and the solid arrows represent reactions and translocations. The 'Regulator' is a transcription factor which is only active after binding the 'Autoinducer' small molecule. Synthesis of the autoinducer is catalyzed by the 'Synthase' protein. Amongst the target genes upregulated by the active transcription factor is that of the synthesizer protein. This positive feedback loop is thought to drive the system strongly from the uninduced to the induced state.
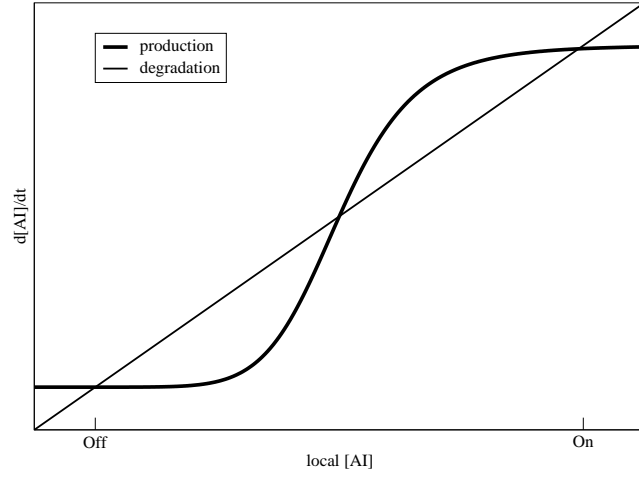
Figure 2.2:   Schematic of quorum sensing dynamics. The simple, coarse grained picture of quorum sensing as a balance between linear degradation and sigmoidal production rates for the autoinducer as a function of the local concentration. The steady state solutions of this model occur at the intersections of the two curves, where production is matched by degradation. In this model there are two stable steady state solutions, labeled 'Off' and 'On', and one unstable solution at an intermediate level. Stochastic fluctuations of the local autoinducer concentration across the unstable steady state level drives the system to the other state.

(e) a quorum is reached, the bacteria (f) duplicate an extrachromosomal ring of DNA called the Tumor inducing (Ti) plasmid that contains all the genes necessary for pathogenesis and (g) inject the new copies into neighboring recipient bacteria. Finally, (h) a number of signals cause the bacteria to mitigate the active quorum response. The computational model we have developed, and detail in Section 2.3, encompasses primarily steps (c) through (f).

Plant cells, especially at sites of wounding, release a class of chemicals called phenolic compounds (e.g. acetosyringone). These phenolic compounds serve as the chemical signal to initiate an infection. They are (a) sensed by the VirA-VirG two component regulatory system which then induces the virulence (Vir) network. Induction of virulence leads to (b) the transfer of oncogenic DNA from a portion of the Ti plasmid into the plant cell. Once this transfered DNA (T-DNA) is incorporated in the plant genome, the plant cells express the genes encoded for in this DNA sequence. Some of the genes encoded for on the T-DNA cause the plant cells to synthesize growth hormones, leading the local plant cell population to proliferate and form tumors. Other genes on the T-DNA cause the plant to synthesize and release a class of chemicals called opines (e.g. octopine) which the bacteria use as a carbon, nitrogen, and energy source.

From this point the process is also depicted in Figure 2.3. Octopine is the signal that the infection is successful and the bacteria can begin to look for a quorum. Octopine taken up by the bacterial cells (c) activates transcription (by relieving repression) of the *traR* gene (which codes for the **Regulator** in Figure 2.1). This derepression is accomplished by binding to the OccR protein complex (a tetramer of OccR already bound to DNA upstream of the *traR* gene) and relieving a high-angle DNA bend [1]. In Figure 2.3 this is depicted in the upper left corner. Low

basal expression of TraI (the **Synthase** in Figure 2.1) meanwhile leads to correspondingly low basal synthesis of *N*-3-oxooctanoyl homoserine lactone (OOHL) (the **Autoinducer** in Figure 2.1). While (d) OOHL levels are low, TraR protein is non-functional because it can not fold into a stable tertiary structure and is degraded [39] (upper middle of Figure 2.3).

Once OOHL levels accumulate (presumably due to a quorum being reached), (e) TraR binds OOHL quickly enough to fold correctly and stabilize [39, 40]. Stable TraR then dimerizes (upper right of Figure 2.3), binds DNA sequences known as tra boxes and recruits RNA polymerase, thereby upregulating transcription of target genes [12] (lower portions of Figure 2.3). One of these tra boxes is upstream of the *traI* gene and is responsible for a large upregulation of TraI expression (the positive feedback loop described in Section 2.1). The other target genes are mostly comprised of the genes necessary for conjugal transfer (hence the tra prefix) of the Ti plasmid and are not depicted in Figure 2.3. Conjugal transfer consists of (f) duplicating the plasmid, expressing and assembling a type IV secretion/mating-pore formation system (a needle-like structure with a larger base in the membrane composed of a variety of proteins) and (g) injecting the duplicated DNA into neighboring bacteria. There is therefore an indirect activation of the *traR* gene due to the increase in gene copy number [26].

At this point (h) a number of modes of negative regulation engage. One of the TraR target genes, *traM*, codes for a negative regulator of TraR: TraM proteins bind TraR and sequester them from the DNA (right center of Figure 2.3). Another negative regulator is TrlR which has high sequence similarity to the N-terminal ligand binding and dimerization portion of TraR without the DNA binding domain. Expression of the *trlR* gene is under control of another opine, mannopine, and

Figure 2.3: Quorum sensing in *Agrobacterium tumefaciens*. A more detailed depiction of the quorum sensing network is depicted here. Colored shapes are proteins, thick horizontal arrows are genes, black circles, triangles and pentagons are small molecules, and X's represent degradation. The bacterial cell is not depicted and neither is plasmid duplication. All genes depicted in this figure are on the Ti plasmid and hence are duplicated and shared in response to sensing a quorum.

the TrlR protein inhibits TraR activity by binding OOHL and also stable TraR forming heterodimers [8] (center of Figure 2.3). The final known type of negative regulation is by AttM, a protein with OOHL degrading activity [37] (not depicted in Figure 2.3). The primary substrate of AttM appears to be a separate class of molecules not connected to quorum sensing however and the rate at which it degrades OOHL is quite low.

2.3   Computational Model

We developed a computational model of the quorum sensing network in *Agrobacterium tumefaciens*. Our model of this network consists of the following simpli-

fications of the description in Section 2.2: OccR is not specifically in the model so initial activation of *traR* is directly by octopine. Everything related to TrlR is absent (that is, the TrlR protein, mRNA and gene as well as mannopine). A version of the model is under development which includes TraM but this implementation is incomplete; the version of the model described here therefore is without TraM. The activation of gene expression by stable TraR dimers is implemented as a Michaelis-Menten type reaction since the transcription factor:DNA complex is thought to be in equilibrium. The equations defined by this network are available in Appendix E.

The model encompasses twenty-seven biochemical interactions involving twenty different molecular species (e.g. genes, mRNA, and proteins in different oligomeric states). This reaction network is described as a coupled system of first order ordinary differential equations giving the time rate of change of each species' concentration. This model contains twenty four unknown parameters—rate and Michaelis Menten constants—which we do not know but which are necessary to make any quantitative predictions. In order to estimate these parameter values, data has been collected from the literature and the parameters have been optimized to find a best fit. This is accomplished by defining a cost function that quantifies how different the model output is for a given set of parameters from the experimental data. The precise function is a sum of squared differences between the model and the data, scaled by the experimental error:

$$C(\mathbf{p}) = \sum_{i=1}^{N} \frac{(y_i(\mathbf{p}) - d_i)^2}{\sigma_i^2} \tag{2.1}$$

where d_i is the ith data point, σ_i its associated error and $y_i(\mathbf{p})$ is the output of the model for the same measurement (e.g. the concentration of a particular protein at a particular time for a given set of initial conditions). Up to overall normalization

factors, this is a χ^2 measure of the parameter fit.

If we assume that the experimental errors are uncorrelated, random, and Gaussian, the relative probability that the model with a given set of parameters would produce the observed data is

$$P(D|\mathbf{p}) \propto \prod_{i=1}^{N} \exp\left(\frac{-(y_i(\mathbf{p}) - d_i)^2}{\sigma_i^2}\right) \tag{2.2}$$

where D is the set of all datapoints $\{d_i, \sigma_i\}$. Transforming the product over exponentials into a sum over their exponents, we see that $P(D|\mathbf{p}) \propto \exp(-C(\mathbf{p}))$ and that the set of parameters that minimizes the cost function also maximizes the probability. We can then use Bayes theorem to find the relative probability of a particular set of parameters given the data:

$$P(\mathbf{p}|D) \propto P(D|\mathbf{p})P(\mathbf{p}). \tag{2.3}$$

Here, $P(\mathbf{p})$ is the prior probability for the parameters. In the optimization process that was conducted for this model, a uniform probability was used for all parameter values so the term $P(\mathbf{p})$ is simply incorporated into the proportionality. This uninformative prior avoids possibly incorrect biasing of the parameter estimates but it could be replaced by a sufficiently weak distribution to avoid unphysical parameter values.

This optimization procedure is not trivial: when the parameters are substantially far from producing the correct dynamics then no small change in parameter values affects the fit and hence no gradient information is available to guide the search. Different directions in parameter space also tend to have widely different natural scales (i.e. the amount of change in a given direction required to achieve a given change in the cost function) so the optimization algorithm must be able to gauge these differences and alter step sizes accordingly. For this reason the

Nelder-Mead simplex algorithm [28] was used to begin the optimization and once it converged, Levenberg-Marquardt [28] was used to find a more precise minimum. The best fit parameters are available in Appendix F as well as the fits to the data.

While the best fit set of parameters do have a low cost (and do fit the training data well) they should not be quoted as 'true' values unless the estimates are very tightly constrained about this best fit. To determine how well constrained the parameters are in the vicinity of the best fit we calculate the second derivative of the cost with respect to the parameter values. This matrix is called the Hessian:

$$H_{ij} = \frac{\partial^2 C}{\partial \log(p_i) \partial \log(p_j)}. \tag{2.4}$$

The derivatives are taken with respect to the logarithms of the parameters because different parameters may have different units and this form calculates the change in the cost function for a given *fold* change in the parameter values. For cost functions such as Equation 2.1 that are a sum of squares we can approximate the Hessian by noting that for a good fit to the data, each residual must be small. Expanding the second derivative:

$$\frac{\partial^2 C}{\partial \log(p_i) \partial \log(p_j)} = \sum_{k=1}^{N} \left(\frac{\partial r_k}{\partial \log(p_i)} \frac{\partial r_k}{\partial \log(p_j)} + r_k \frac{\partial^2 r_k}{\partial \log(p_i) \partial \log(p_j)} \right) \tag{2.5}$$

we see that the second term can be dropped in the case of a near-perfect fit when each r_k is small. Denoting the matrix of first derivatives as the Jacobian,

$$J_{kj} = \frac{\partial r_k}{\partial \log(p_j)} \tag{2.6}$$

we can then make the approximation $H \approx J^\top J$. In the Bayesian statistics field this matrix is known as the Fisher Information Matrix and the Jacobian is referred to as the Design Matrix for the linearized approximation of the full model.

The Hessian defines the quadratic approximation to the cost function surface about the best fit parameters. If we take the eigenvalue decomposition of the Hessian we can characterize the ellipses defined by this approximation; the principal axes of the ellipse are the eigenvectors and the curvature along each eigenvector is given by the corresponding eigenvalue. If a given direction in parameter space has high (low) curvature then the cost function is a quickly (slowly) rising function of that combination of parameters. If the cost function is rising quickly then there is very low uncertainty in that combination of parameters but if the curvature is very low then large changes in that combination of parameters lead to minimal changes in the cost function and that combination of parameters is very unconstrained. More precisely, the uncertainty in the combination of parameters defined by the ith eigenvector is given by $1/\sqrt{\lambda_i}$ where λ_i is the ith eigenvalue. Figure 2.4 shows the eigenvalues of $J^\top J$ about the best fit. Notice the fantastically large range of eigenvalues— contours of constant cost about the best fit are ellipses with aspect ratios of $\sqrt{\lambda_1/\lambda_N} \approx 10^{28}$.

Since nearly half of the total range of eigenvalues is covered by just the smallest three eigenvalues, it is worth examining the composition of these eigenvectors (available in Appendix F). Seven of the nine eigenvectors with smallest eigenvalue are each dominated by a single bare parameter, suggesting that the model could be redefined to remove these parameters since, at least in the quadratic approximation about this point in parameter space and with this set of training data, changes in their values have little to no influence on the model behavior. This does not mean that those parameters can blindly be set to zero. Depending on their role in the model, it may be appropriate to set some to zero, some to infinity and others simply to fixed values that are not free to change. The other two eigenvectors at
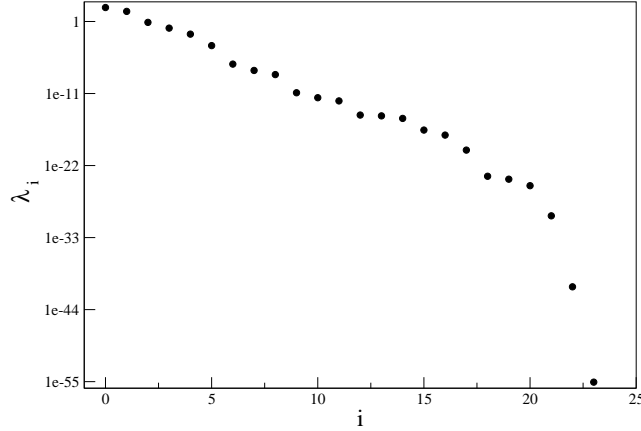
Figure 2.4: Eigenvalues of quorum sensing model Hessian. The eigenvalues of the approximate Hessian ($J^\top J$) about the best fit parameters are plotted on semilog axes.

this low end of the spectrum are the product and ratio of a pair of bare parameters, suggesting that they too could be removed from the network. Even if these nine parameters can be successfully removed from the model, the remaining fifteen parameters still have eigenvalues spanning sixteen orders of magnitude (cost contour ellipses with aspect ratios of 10^8). The composition of these eigenvectors is not well-localized so further model reduction would need to be very sophisticated.

Models with sensitivity spectra such as that in Figure 2.4 (roughly equal spacings between the logarithms of eigenvalues, causing an exponentially large total range) are by definition *sloppy* [4, 5]. Such models have many free parameters but the model behavior is dominated by only a small number of stiff, particular combinations of parameters. These dominant parameter combinations correspond to the eigenvectors of the Hessian with large eigenvalues. Because sloppy models are so insensitive to moves in the majority of parameter space (i.e. any move that is not along a stiff direction), the true parameter values can never be constrained

by the collective behavior of the model [16]. The general condition of sloppiness, its origins, and universality are developed in Chapter 1.

2.4   Future Work

While the current model does fit the current set of data there is much that could be done to improve it. The first step would be to build an ensemble of parameter sets proportional to how well they fit the data. This is necessary for making any predictions about future experiments (or past experiments that were not in the training set). An ensemble is necessary because a) the best fit parameters are not well constrained so any prediction needs to be accompanied by error bars representing the parameter uncertainty and b) the nonlinear mapping from parameter values to model behavior is fully captured by an ensemble but not by simple propagation of the quadratic estimates about the best fit. Such an ensemble is built using Markov Chain Monte Carlo (MCMC) techniques. A chain of points in parameter space is assembled such that the distribution of these points matches the probability distribution defined by the cost function. Sampling from points in this chain is then a reliable substitute for sampling from the original distribution. The wide range of natural scales for different directions in parameter space (reflected in the wide range of eigenvalues of the Hessian) makes constructing this chain a nontrivial problem. Step sizes too large along stiff directions will never be accepted but steps too short along sloppy directions will never converge. A Metropolis-Hastings algorithm, which uses the Hessian to guide step sizes in different direcetion, is therefore necessary for achieving convergence [7].

Another improvement is that more data could be added to the training set.

Quorum sensing in *Agrobacterium* has been under investigation by a number of labs for many years and more data is still available. More data might couple the nine parameters dominating the sloppiest eigenmodes to the rest of the system. Several known components of the quorum sensing network have been left out of the current model (e.g. negative regulation by TraR sequestration with TrlR or TraM and OOHL degradation by AttM) because no data related to those mechanisms were included in the training set. Incorporating more data would allow for expanding the definition of the model to include these mechanisms. Before any further optimization is done (or in fact before an ensemble is generated), more informative priors should be placed on the parameter values. In other models of biological networks it has been found that the system dynamics alone still allow for parameter values that are plainly not relevant or physically possible [4, 16]. In these situations, any educated limits on the parameter values can be of great aid in hastening the optimization procedure and preventing the ensembles from encountering a variety of computational difficulties arising from unphysical parameter values.

In the area of changes to model structure, it should be noted that most of the relevant experiments probe the structure of this network more than the quantitative dynamics. Experiments that knock out a given gene and discover that quorum sensing is completely blocked (e.g. Figure F.2) are extraordinarily useful in answering the question of whether the gene (or protein product) is part of the network but offer little constraint on any biochemical parameters. For this reason, it would be worthwhile considering whether a different structure entirely for the model might be more useful. Bayesian networks, which describe the network of interactions simply by conditional probabilities (if X is high then, with probability

p, Y is low) may connect to this type of data more closely then the differential equations describing the biomolecular interactions. Another interesting question about the type of model being developed is the absence of spatial structure. The dimensionality of the experimental setup (i.e. whether the bacteria are cultured in 3D broth culture or on 2D plates) has been shown to have significant effects on the process of quorum sensing [11] but the current form of the model incorporates no spatial information. The effects of dimensionality could be due to active signaling processes but it would be interesting to learn whether the differences between 2D and 3D diffusion of the autoinducer could explain the differences.

One very interesting question about the quorum sensing network in *Agrobacterium tumefaciens* that the model might be useful in understanding is the effects of noise on the activation of the network. Before the quorum is counted, the levels of TraR in the cell are so low that the difference between a few proteins and no proteins might be relevant. The positive feedback loop should also amplify the effects of this noise—if just enough copies of TraR happen to bind OOHL in one bacterium but not another, the expression of TraI will be greatly upregulated in the former relative to the latter. This could lead to great differences in the activation states of different bacteria under similar experimental conditions. Preliminary experiments in the lab of Dr. Winans suggest that this is indeed the case, where GFP under the control of a promoter responsive to TraR shows noisy induction kinetics from culture to culture [35]. The model could perhaps be used to suggest which components of the network have particularly strong influences on this stochastic onset and what experiments could reveal this effect.

Appendix A

Eigenvalues of *ESE*

Figures A.1 and A.2 are histograms demonstrating that the eigenvalues of matrices of the form ESE are bounded by the corresponding row sums of EES (Conjecture 1). In creating such histograms there are a few choices one can make: the size of the matrices, the distribution for the elements of $\tilde{S}$ ($S = \tilde{S}^\top \tilde{S}$), and the value of ϵ (it must be between $0 < \epsilon \le 1$. Empirically we find that smaller system sizes lead to eigenvalues which approach the row sum bound more closely so Figures A.1 and A.2 are for 2×2 matrices. Larger values of ϵ also lead to ratios λ_i/r_i closer to one. We find that if the absolute value of the mean of the distribution for $\tilde{S}_{ij}$ is large, then λ_1 is closer to r_1 ($\lambda_1/r_1 \approx 1$) but λ_n is further from r_n ($\lambda_n/r_n \approx 0$). Neither the sign of the mean nor the width of the distribution of $\tilde{S}_{ij}$ appears to have any significant effect on the relationship between the eigenvalues and the row sums. In general we find that for a given system size, distribution for $\tilde{S}_{ij}$, and ϵ, the larger eigenvalues approach the row sum bound much closer than smaller eigenvalues. This is demonstrated by the fact that the histogram in Figure A.1 is dominated by ratios near one while Figure A.2 is dominated by ratios near zero.

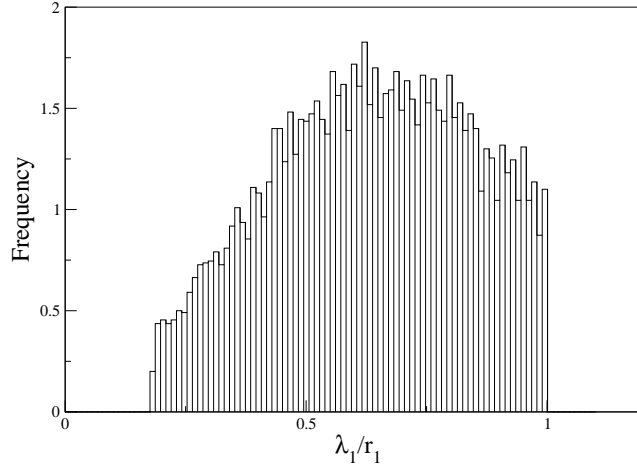We were lead to this discovery by considering the Gershgorin Circle Theo-

Figure A.1: Row sum bound on first eigenvalue of ESE. Histogram of ratios of largest eigenvalue, λ_1, of ESE to first row sum of EES, $r_1 = \epsilon^0 \sum_i |S_{1i}|$. For this ensemble both E and S are 2×2 matrices and $\epsilon = 1/10$. S is formed by creating a 2×2 matrix, $\tilde{S}$ with each element selected randomly from a Gaussian distribution with mean 0 and standard deviation 1 ($\tilde{S}_{ij} = N(0,1)$) and then forming the symmetric, positive definite matrix S by $S = \tilde{S}^{\top}\tilde{S}$. Note the hard wall at $\lambda_1/r_1 = 1$ showing that over the entire ensemble λ_1 came very close to, but never superseded, r_1.
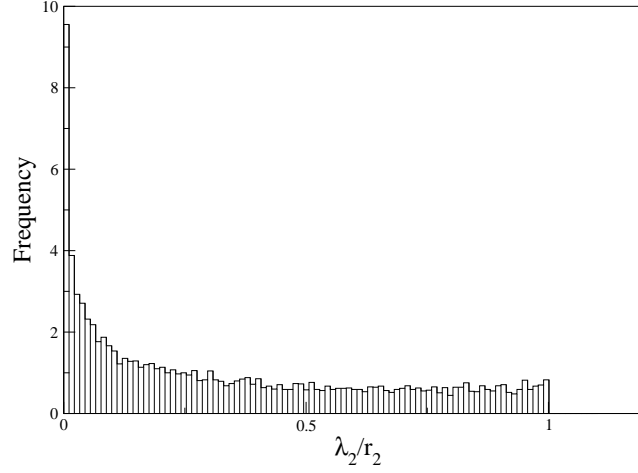
Figure A.2: Row sum bound on second eigenvalue of ESE. Histogram of ratios of second eigenvalue, λ_2, of ESE to second row sum of EES, $r_2 = \epsilon^2 \sum_i |S_{2i}|$. For this ensemble both E and S are 2×2 matrices and $\epsilon = 1/10$. S is formed by creating a 2×2 matrix, $\tilde{S}$ with each element selected randomly from a Gaussian distribution with mean 0 and standard deviation 1 ($\tilde{S}_{ij} = \mathrm{N}(0,1)$) and then forming the symmetric, positive definite matrix S by $S = \tilde{S}^\top \tilde{S}$. Note the hard wall at $\lambda_1/r_1 = 1$ showing that over the entire ensemble λ_1 came very close to but never superseded r_1.

rem [14]. While the original theorem pertains to the eigenvalues of complex, unsymmetric matrices we quote here the more relevant result for real, symmetric matrices:

Theorem 1 *Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric and that $Q \in \mathbb{R}^{n \times n}$ is orthogonal. If $Q^\top A Q = D + F$ where $D = \mathrm{diag}(d_1, \ldots, d_n)$ and F has zero diagonal entries, then*

$$\lambda(A) \subseteq \bigcup_{i=1}^{n} [d_i - r_i, d_i + r_i] \tag{A.1}$$

where $r_i = \sum_{j=1}^{n} |f_{ij}|$ for $i = 1 : n$.

This result provides a very useful bound for the eigenvalues of a matrix: the eigenvalues of A are contained within the n Gershgorin Circles of radius r_i centered at d_i. Note that there is no guarantee that each circle contains an eigenvalue, simply that the space covered by all the circles contains all the eigenvalues. Considering the off diagonal elements, F as a perturbation on D and considering how the eigenvalues of $D + F$ move as the perturbation becomes larger, it is clear that each disconnected set of Gershgorin circles contains precisely as many eigenvalues as their are overlapping circles. Concretely, if a 4×4 system has three overlapping circles and one separate circle, then the one separated circle contains precisely one eigenvalue while the union of the three other cirlces contains the other three eigenvalues. It may easily be the case that of these three overlapping circles, one or two do not contain any eigenvalues however. A similar situation is depicted in Figure A.3—one Gershgorin circle is wholly contained within the other but contains no eigenvalues itself. The row sum bound we describe in Conjecture 1 is a slight modification of the Gershgorin Circle Theorem where we use the similarity transform $E(ESE)E^{-1} = EES$ and can now place (at least) one eigenvalue within
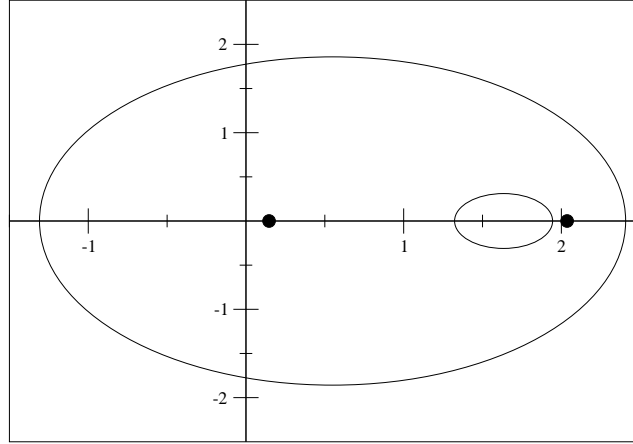
70

Figure A.3: Gershgorin circle bounds on eigenvalues. The eigenvalues of a symmetric, positive definite 2×2 matrix and the corresponding Gershgorin Circle bounds ($Q = I$, the Identity matrix, in Theorem 1). Note that since the two Gershgorin circles overlap, the theorem allows one circle to contain no eigenvalues.

each circle but they are instead centered at the origin with a radius given by $|d_i| + r_i$ and A must be positive definite. This bound is depicted in Figure A.4 which is the same system (same A, Q, D, and F) as for Figure A.3.
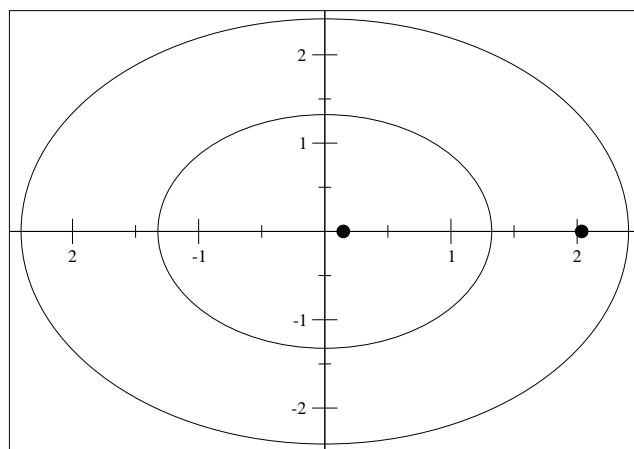
Figure A.4:   Row sum bounds on eigenvalues. The eigenvalues of a symmetric, positive definite 2×2 matrix and the corresponding row sum bounds from Conjecture 1. Note that, as opposed to Figure A.3, each circle contains at least one eigenvalue.

Appendix B

Eigenvectors of Vandermonde Ensemble

Figures B.1 and B.2 are histograms of the dot products between eigenvectors of random Hessians ($H = V^\top A^\top AV$) and the corresponding right singular vectors of the Vandermonde matrices ($V_{ij} = \epsilon_j^{(i-1)}$). The difference between the two figures is that the typical size of ϵ in Figure B.1 is ten thousand times larger than in Figure B.2. Even over this large range, the eigenvalues are incredibly well-aligned with the right singular vectors, since even for the larger ϵ the overwhelming majority of dot products are near one. Furthermore, the fact that the width of these distributions is correlated with the size of ϵ is evidence that as ϵ decreases the alignment improves. Since both eigenvectors and singular vectors have unit length, a dot product of one means that the angle between the two vectors is near zero, ($\vec{a} \cdot \vec{b} = |a||b| \cos(\theta)$ for any two vectors $\vec{a}$ and $\vec{b}$).

Depicted in Figures B.3 and B.4 are the eigenvectors of the Hessian after it has been transformed into the basis of right singular vectors of the Vandermonde matrix. If this basis were exactly the eigenvectors of the original untransformed

Hessian, these vectors would be standard basis vectors (the kth standard basis vector has value one at index k and zero at all other indices). Figure B.3 depicts an ensemble of size five hundred for fixed eigenvector number (three) and fixed ϵ (1/1000). The various members of the ensemble are plotted as circles and the function $y = \epsilon^{|x-3|}$ (here $\epsilon = 1/1000$) is shown to make the scaling behavior of the eigenvector components clearer. Figure B.4 differs only in that a different eigenvector is plotted (the fifth) and a different value of ϵ (1/100) is used to demonstrate that the results are general. It is clear that while there is an interesting structure to the higher order corrections, they do approach zero as $\epsilon \to 0$ and thus, to leading order in ϵ the eigenvectors of the Hessian are indeed the right singular vectors of the Vandermonde matrix.
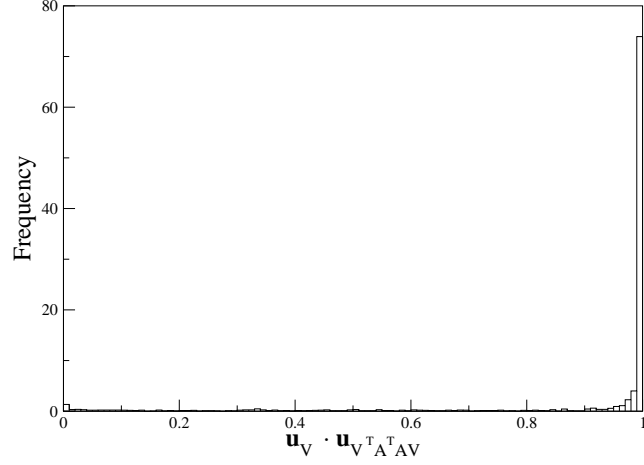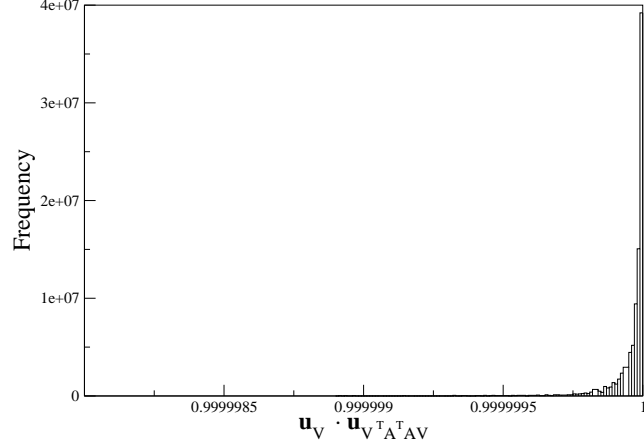
Figure B.1: Eigenvectors of Vandermonde ensemble matrices with large ϵ. Histogram of dot products between eigenvectors of random Hessians $(H = V^\top A^\top A V)$ and right singular vectors of Vandermonde matrices (V). An ensemble of size 500 was generated where each matrix had dimensions 6×6, the elements of A were drawn from a normal distribution with mean 0 and variance 1 $(A_{ij} = \mathrm{N}(0, 1))$, and 6 'parameters' defining the Vandermonde matrix $(V_{ij} = \epsilon_j^{(i-1)})$ were selected from a log uniform distribution between -10 and 10 $(\ln(\epsilon_j) = \mathrm{U}(-10, 10))$.
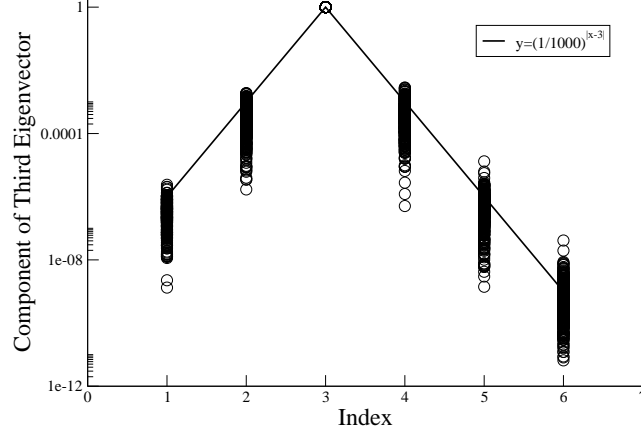
Figure B.2: Eigenvectors of Vandermonde ensemble matrices with small ϵ. Histogram of dot products between eigenvectors of random Hessians ($H = V^\top A^\top A V$) and right singular vectors of Vandermonde matrices (V). Note the remarkably small range for the horizontal axes. An ensemble of size 500 was generated where each matrix had dimensions 6×6, the elements of A were drawn from a normal distribution with mean zero and variance one ($A_{ij} = \mathrm{N}(0,1)$), and 6 'parameters' defining the Vandermonde matrix ($V_{ij} = \epsilon_j^{(i-1)}$) were selected from a log uniform distribution between -1/100 and 1/100 ($\ln(\epsilon_j) = \mathrm{U}(-1/100, 1/100)$).

Figure B.3: Third eigenvector components in Vandermonde ensemble. Ensemble of components of third eigenvector for Hessians ($H = V^\top A^\top A V$) transformed into basis of right eigenvectors of the Vandermonde matrix (V), $\tilde{H}$ (Conclusion 3). The matrices are all of size 6×6. $\tilde{H} = \Sigma^\top A^\top A \Sigma$ was generated by creating a random matrix A with elements from a normal distribution with mean zero and variance one ($A_{ij} = N(0,1)$) and a diagonal matrix $\Sigma_{ii} = \epsilon^{(i-1)}$ with $\epsilon = 1/1000$. If the eigenvectors of H were the same as those of V, then $\tilde{H}$ would be diagonal and its eigenvectors would be the standard basis vectors, $e_i^{(j)} = \delta_{ij}$ (the jth standard basis vector has value one at index j and zero everywhere else). As can be seen from the plot, the eigenvectors of $\tilde{H}$ are indeed the standard basis vectors to lowest order. The corrections for the jth eigenvector are $\mathcal{O}(\epsilon^{|x-j|})$. The absolute value of all eigenvector components is plotted here to clarify the scaling behavior.
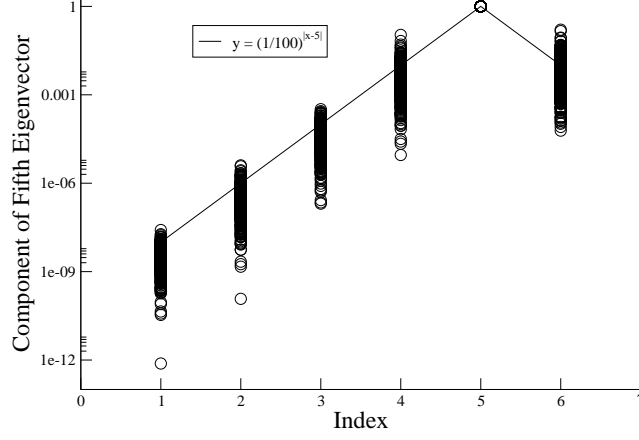
Figure B.4: Fifth eigenvector components in Vandermonde ensemble. Ensemble of components of fifth eigenvector for Hessians ($H = V^\top A^\top A V$) transformed into basis of right eigenvectors of the Vandermonde matrix (V), $\tilde{H}$ (Conclusion 3). The matrices are all of size 6×6. $\tilde{H} = \Sigma^\top A^\top A \Sigma$ was generated by creating a random matrix A with elements from a normal distribution with mean zero and variance one ($A_{ij} = N(0,1)$) and a diagonal matrix $\Sigma_{ii} = \epsilon^{(i-1)}$ with $\epsilon = 1/100$. If the eigenvectors of H were the same as those of V, then $\tilde{H}$ would be diagonal and its eigenvectors would be the standard basis vectors, $e_i^{(j)} = \delta_{ij}$ (the jth standard basis vector has value one at index j and zero everywhere else). As can be seen from the plot, the eigenvectors of $\tilde{H}$ are indeed the standard basis vectors to lowest order. The corrections for the jth eigenvector are $\mathcal{O}(\epsilon^{|x-j|})$. The absolute value of all eigenvector components is plotted here to clarify the scaling behavior.

Appendix C

Subsystem Sloppiness

How many parameters should a model have before we expect it to look sloppy? One convenient method we can use is to study subsystems of larger models. Consider taking the 48 parameter PC12 growth factor model described in Section 1.2, fixing thirty eight of those parameters and allowing the remaining ten to be free. The Hessian for this restricted model would simply be the 10×10 submatrix of the original Hessian defined by these parameters. Since there are $\binom{48}{10} \approx 6.5$ billion different ten parameter submodels it is then quite easy to get good statistics on their sloppiness. In Figure C.1 we assemble ensembles of five-, eight-, ten-, and twelve-parameter submodels from the PC12 network and calculate the eigenvalues of their Hessians. A typical five parameter submodel may or may not be sloppy, quite a few have eigenvalues that only span one or two orders of magnitude. A typical eight parameter model is, however, certainly sloppy with typical eigenvalue ranges of five orders of magnitude (less than two parameters per decade). Not only is the total range of eigenvalues large, they are spaced roughly equally in logarithms.

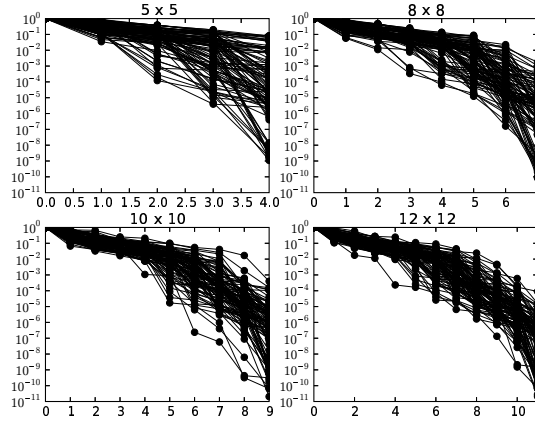From Figure C.1 we have strong evidence that the eigenvalue distribution for

Figure C.1: The eigenvalues of subsystems of the PC12 model with varying numbers of parameters. Notice that the typical five parameter submodel is arguably sloppy and that the typical eight parameter is certainly so.

relatively small models is recognizably sloppy. Now we would like to know the statistics of sloppiness for a fixed number of parameters. In Figure C.2 we take 10-parameter subsystems of three separate biological models ((a) the PC12 network described in Section 1.2, (b) a model of EGF receptor signaling, trafficking, and down-regulation [6], and (c) a model of the yeast cell-cycle [9]) and examine the total range of eigenvalues in these subsystems. The total range of eigenvalues, $\lambda_{max}/\lambda_{min}$, is a quantity called the *condition number*. In each case we see that, over the entire ensemble, all 10-parameter models have strikingly large ranges of eigenvalues, even if only the first of the peaks is considered. We conclude that real-world models with more than about ten parameters are likely to be sloppy.

The remainder of this discussion will be concerned with the clear multi-modal structure in these plots. In brief, we find that there exist individual parameters as well as small groups of parameters that are particularly unconstrained. If they
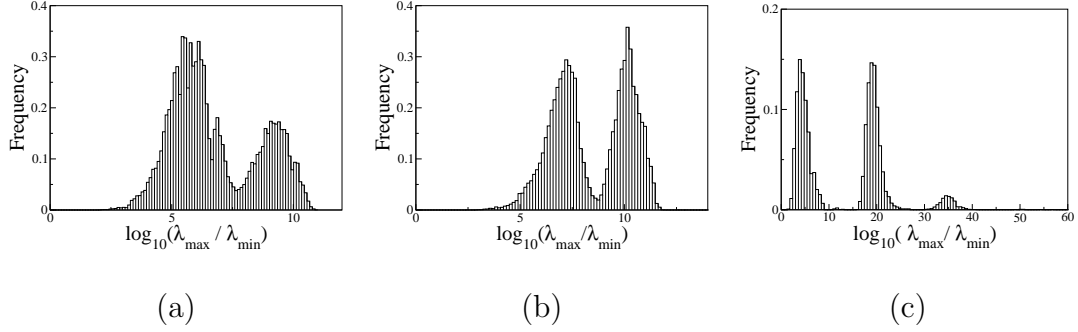
Figure C.2:   Condition number for subsystems of biological models. The condition number ($\lambda_{max}/\lambda_{min}$) for ten parameter subsystems of (a) the PC12 model [4, 16], (b) the EGFR model [6], and (c) the yeast cell-cycle model [9]. The multimodal structures are due to the presence or absence of sets of particularly unconstrained parameters.

happen to be included in a given 10-parameter submodel, then the condition number is particularly large. This kind of few-parameter model degeneracy is what we naïvely expected to find, and we view it as distinct from the collective, emergent sloppiness on which we focus elsewhere.

In Figure C.2 (a) we see that when considering just ten parameter submodels of the PC12 network and looking at only the total range of the eigenvalues there appear to be two separate classes. A given submodel will have a (log base ten) condition number from one of two Gaussian distributions, a lower one centered at roughly six and a higher one centered at roughly nine.

By analyzing the frequency with which the original 48 parameters appears in submodels from these two classes we see that the distinction arises from a few particularly unconstrained parameters or sets of parameters. Figure C.3 (a) shows the frequency with which each of the 48 parameters appears in a submodel in the lower condition number class and Figure C.3 (b) shows the frequency for the higher
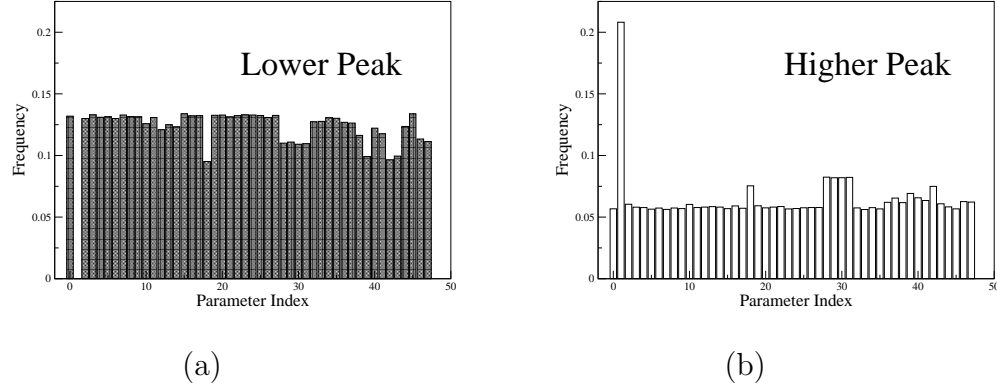
Figure C.3: PC12 submodel parameter frequencies. Each plot is the frequency with which each of the 48 parameters in the PC12 model appear in a 10-parameter subsystem. The two distributions are for (a) the low condition number peak ($\log_{10}(\lambda_{max}/\lambda_{min}) < 6.6$) and (b) the high condition number peak ($\log_{10}(\lambda_{max}/\lambda_{min}) > 7.6$) of Figure C.2 (a).

class. The most striking difference is that parameter one never appears in the lower class. This is the rate constant for unbinding of EGF and the EGF Receptor. This reaction was separately identified by other means as being unnecessary—in all the experimental conditions considered with the model all of the receptors become bound (EGF is either absent or in excess) and the unbinding rate has no effect on the fits, as long as it is small enough to be effectively zero.

The other effects that lead to particularly high condition numbers are slightly more subtle. By comparing Figure C.3 (a) and (b) we see that parameters 18, 42, and 28 through 31 are noticeably enriched in the higher peak and depleted in the lower peak. These six parameters fall into two sets. Parameters 18 and 42 are both involved in the protein BRaf (its activation by Rap1 and its activation of Mek1/2) while parameters 28 through 31 are involved in the PI3K branch of the network. As opposed to the situation with the EGF/EGFR unbinding rate, these parameters do not necessarily lead to large condition numbers just by themselves.
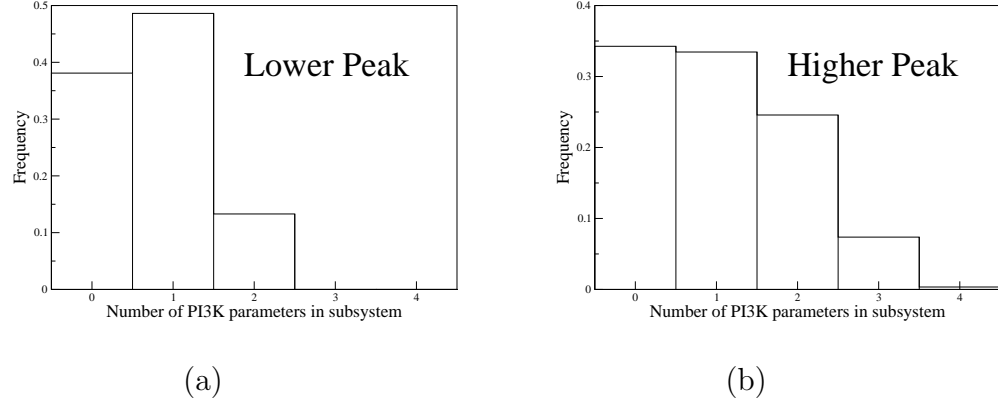
82

(a)   (b)

Figure C.4:   Cooccurrence of PI3K parameters in PC12 submodels. Each plot is the frequency with which the four PI3K parameters from the PC12 network occur together in 10-parameter submodels. The two distributions are for (a) the low condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) < 6.6)$ and (b) the high condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) > 7.6)$ of Figure C.2 (a).

In Figure C.4 we show the frequency with which the four PI3K related parameters occur never, alone, as a pair, a triple, or all together in (a) the lower condition number peak and (b) the higher condition number peak. The higher condition number peak is clearly enriched for cooccurrences of these parameters, providing another source of redundancy expanding the range of the eigenvalues.

In Figures C.5 (a) and (b) we show similar plots for the pair of BRaf related parameters and it is clear that when these parameters occur together, the (sub)model is particularly ill-conditioned.

We have performed similar analyses of two other models and in both instances we get similar results. The first of these models is of EGF Receptor signaling, trafficking, and down-regulation [6]. Figure C.2 (b) shows the distribution of (log base 10) condition numbers for ten parameter submodels of this EGFR model. Again there is a clear double-peak structure. By examining the relative frequencies
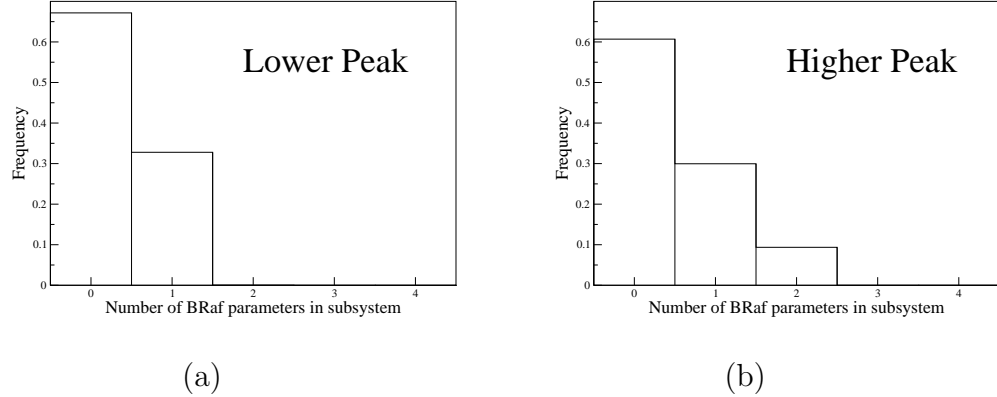
Figure C.5: Cooccurrence of BRaf parameters in PC12 submodels. Each plot is the frequency with which the two BRaf parameters from the PC12 network occur together in 10-parameter submodels. The two distributions are for (a) the low condition number peak ($\log_{10}(\lambda_{max}/\lambda_{min}) < 6.6$) and (b) the high condition number peak ($\log_{10}(\lambda_{max}/\lambda_{min}) > 7.6$) of Figure C.2 (a).

with which each of the 56 original parameters occur in these two peaks we see that the sources of ill-conditioning are even more straight forward than in the PC12 model. Figure C.6 (a) and (b) show the relative frequencies for each of the original 56 parameters in submodels from the lower and higher condition number peak, respectively. Three parameters (indices 12, 44, and 46) stand out as never occurring in the lower peak and being more than twice as common as the other 53 parameters in the higher peak. While not every single submodel in the higher peak contains one of these three parameters, the possibility that they are the dominant source of the difference is supported by by the relative number of models in each peak. If the difference between the two is only these three parameters, then the fraction of models in the lower peak should be given by the product of probabilities that when selecting each of the ten parameters, neither of the three were selected. This probability is $\prod_{i=0}^{9} \frac{56-3-i}{56-i} \approx 0.55$ and the actual fraction of the ensemble in
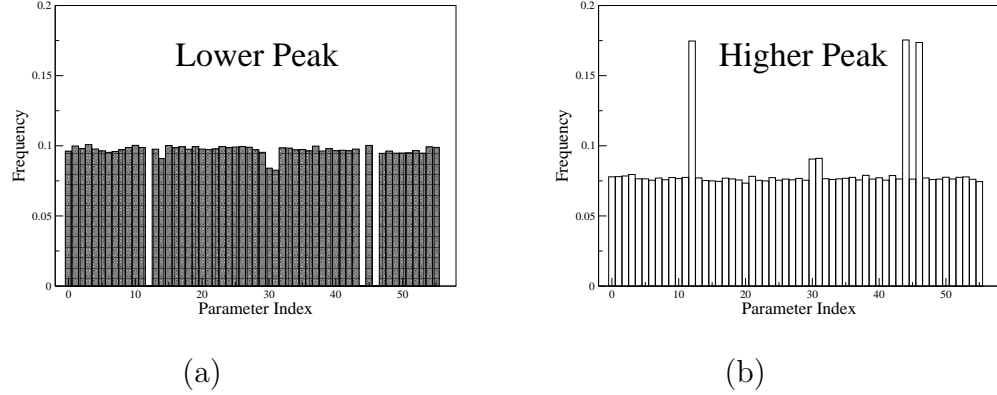
Figure C.6: EGFR submodel parameter frequencies. Each plot is the frequency with which each of the 56 parameters in the EGFR model appear in a 10-parameter subsystem. The two distributions are for (a) the low condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) < 8.2)$ and (b) the high condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) > 9.)$ of Figure C.2 (b).

this peak is about 51% (51,406 out of 100,000 models had a condition number less than $10^{8.2}$). The three parameters that segregate so strongly are (a) the unbinding of the protein Cbl from the heterodimer of proteins Cool-1 and Cdc42, (b) the transcription rate for the EGF Receptor, and (c) the rate at which the Recycling pathway operates. This result confirms other tests indicating that the precise value of these parameters were insignificant for the behavior of the model.

The final model which we have analyzed in this way is a 143-parameter model of the cell-cycle in budding yeast [9]. Figure C.2 (c) shows that the condition numbers for ten parameter submodels fall into three well-separated peaks. By analyzing the relative frequency of the 143 original parameters in each of these peaks we see that there are fourteen special parameters which are not present in the lowest peak (Figure C.7 (a)), are moderately enriched for in the middle peak (Figure C.7 (b)), and are greatly enriched for in the highest peak (Figure C.7

(c)). By calculating the probability of choosing ten random parameters from the original 143 and never selecting any of these fourteen parameters ($\prod_{i=0}^{9}(\frac{143-14-i}{143-i}) \approx 0.344$) and comparing to the relative number of submodels in the lowest peak (34,601/100,000 have condition numbers less than 10^{10}) we see that the cause for these distinct peaks is primarily whether or not any of these parameters is selected. By analyzing the frequency with which these fourteen parameters occur alone, as a pair, as a triple, etc in the middle and highest peaks (Figure C.8 (a) and (b), respectively) we see that, for the most part, if the submodel has one of these parameters then the condition number lands in the mid-range peak and if it has more than one of these parameters the condition number is in the highest peak. Four of these parameters appear in conservation laws (e.g. without degradation and synthesis, the total concentration of some molecular species is constant) that are already satisfied by the differential equations and hence it is not surprising that they are unimportant for proper functioning of the model. In fact the condition number of submodels with these parameters should be infinity because, up to numerical noise, they are exact 0 modes of the model. These modes are reminiscent of gauge invariances in physics where a physical theory has more detail than occurs in nature. For example, in a spin-glass system if the sign of a spin is changed as well as the sign of all the neighboring spins, the total energy is unchaged. Of the remaining special parameters, eight are involved in defining discontinuous transitions occurring in the dynamics (such 'events' are used to represent the various checkpoints in the cell-cycle: when the concentration of species X crosses some threshold, the ODEs are ignored and concentrations are set by some other rule). It is not clear why parameters associated with these events are so redundant. Anecdotally, many of the events in this model are directly

86

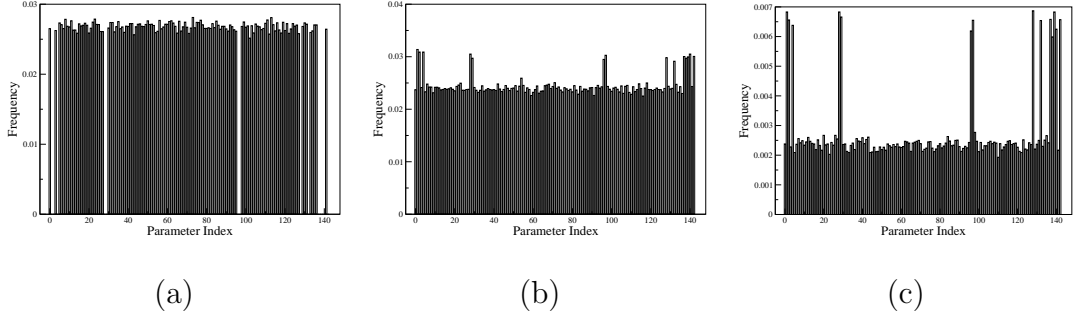| (a) | (b) | (c) |

Figure C.7:   Yeast cell-cycle submodel parameter frequencies. Each plot is the frequency with which each of the 143 parameters in the yeast cell-cycle model appear in a 10-parameter subsystem. The three distributions are for (a) the low condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) < 10)$, (b) the midrange condition number peak $(10 < \log_{10}(\lambda_{max}/\lambda_{min}) < 28)$, and (c) the high condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) > 28)$ of Figure C.2 (c).

triggered by other events [15]. If the events these eight unconstrained parameters appear in are being triggered directly by other events, then no infinitesimal change in the triggering parameters will have any effect on the model behavior. This would explain why at least the trigerring parameters are so unconstrained. The remaining two special parameters are basal synthesis rates for the proteins Cln2 and Pds1. It is again not clear why these basal synthesis rates are so insignificant. Further investigation of this novel analysis could be valuable in understanding the model.
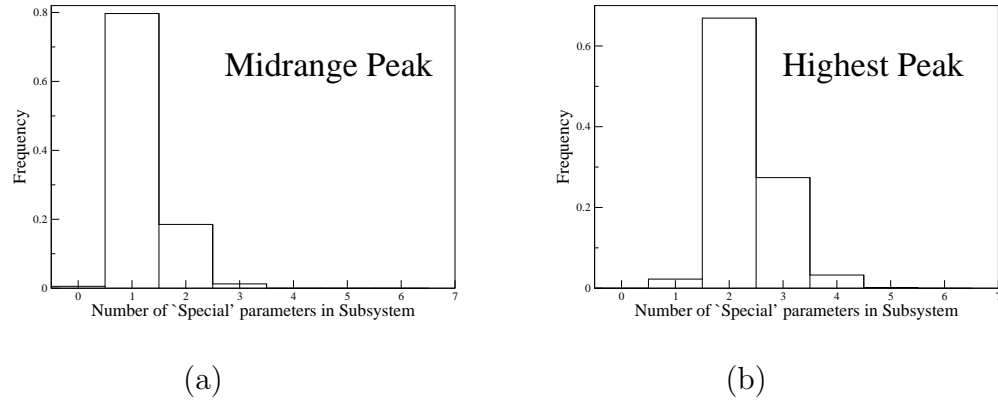
(a)　　　　　　　　　　　　(b)

Figure C.8:　　Cooccurrence of 'special' parameters in yeast cell-cycle submodels. Each plot is the frequency with which the fourteen parameters from the yeast cell-cycle network occur together in 10-parameter submodels. The two distributions are for (a) the midrange condition number peak $(10 < \log_{10}(\lambda_{max}/\lambda_{min}) < 28)$ and (b) the high condition number peak $(\log_{10}(\lambda_{max}/\lambda_{min}) > 28)$ of Figure C.2 (c).

Appendix D

Identifying Sloppy Parameter Sets

Parameters and parameter combinations related to sloppy directions in parameter space are precisely the type of object that should be coarse-grained out to simplify a model. In a true member of the Vandermonde ensemble (Section 1.8) one could remove a degree of freedom simply by reducing the number of parameters by one since the model treats them all symmetrically. Real-world models are not strictly symmetric in each of the parameters, so that procedure becomes infeasible.

One might imagine that removing the sloppiest eigenvectors of the Hessian is the sensible approach since these are by definition the combinations of parameters to which the model is most insensitive. This, however, is not appropriate for the following four reasons. First, the accuracy to which the components of any eigenvector can be resolved is determined by the magnitude of the eigenvalue and its separation from other eigenvalues. The sloppiest eigenvectors are therefore those most affected by noise and while the definition of the large sloppy space is relatively well-determined, the precise composition of the sloppy eigenvectors is not [19]. Second, the sloppy eigenvectors tend not to be well-localized and are instead composed of significant fractions of many of the bare parameters [5]. The

bare parameters are relatively easy to add and remove but combinations of many bare parameters are both difficult to remove from a model sensibly and difficult to learn from (an admirable goal of any coarse-graining process). Third, as described in Section 1.8 and detailed in Appendix B, the composition of eigenvectors for sloppy systems is not usually informative. In that case, it was dominated by by the singular vectors of the Vandermonde matrix, which in turn are determined by the parameters values, not the structure of the model. More broadly, we find strong analogies there to random matrix theory, where eigenvectors in the universal ensembles are uncorrelated and random. Fourth, most real-life models are actually composed of many separate Vandermonde ensemble style systems (Section 1.9). This means that the eigenvectors of the entire system mix together parameters which should be largely uncoupled.

While being able to identify entire subsystems that belong to the Vandermonde ensemble would be incredibly useful, it is a difficult task which we have not solved yet. Instead, we focus here on simply identifying small numbers of bare parameters to which the model is insensitive. Knowing such parameters and pairs of parameters should be useful in simplifying the model.

Let us first identify single parameters which can be most easily removed from a model. As elsewhere in this work we focus on cost functions that are sums-of-squares. The Jacobian, the matrix of first derivatives of the residuals with respect to the parameters, is then $J_{i\alpha} = \partial r_i / \partial p_\alpha$. The Hessian, the matrix of second derivatives of the cost function with respect to the parameters is then

$$H_{\alpha\beta} = \sum_i r_i \frac{\partial^2 r_i}{\partial p_\alpha \partial p_\beta} + \frac{\partial r_i}{\partial p_\alpha} \frac{\partial r_i}{\partial p_\beta} \tag{D.1}$$

$$= \sum_i r_i \frac{\partial^2 r_i}{\partial p_\alpha \partial p_\beta} + \left(J^\top J\right)_{\alpha\beta} \tag{D.2}$$

At the best fit parameters the residuals r_i in general are often small. When this is the case, $r_i \equiv 0$ and we can ignore the first term in equation D.2. This approximation, $H \approx J^\top J$, is the basis of the Levenberg-Marquardt optimization algorithm [28]. Parameter α is ignorable if $J_{i\alpha} \equiv 0$ for all i. The diagonal components of $J^\top J$ are given by the 2-norm of the corresponding columns of the Jacobian $(\left(J^\top J\right)_{\alpha\alpha} = \sum_i J_{i\alpha}^2)$, and so ignorable parameters lead to small diagonal entries of the Hessian. This means that to identify single parameters which can be removed from a model with the least impact, one should identify the columns of the Jacobian with smallest norm.

We now turn to identifying pairs of parameters that lead to sloppiness. Two parameters are 'close' in the model if some linear combination of the two can be removed with minimal impact. Any normalized linear combination of two parameters p_α and p_β can be written as $\tilde{p} = \lambda p_\alpha + \sqrt{1 - \lambda^2} p_\beta$. Motivated by the results in finding single sloppy parameters, we now wish to find α, β, and λ such that $|\partial r_i / \partial \tilde{p}|^2$ is small. Expanding this derivative, we have

$$\frac{\partial r_i}{\partial \tilde{p}} = \lambda \frac{\partial r_i}{\partial p_\alpha} + \sqrt{1 - \lambda^2} \frac{\partial r_i}{\partial p_\beta} \tag{D.3}$$

$$\left|\frac{\partial r_i}{\partial \tilde{p}}\right|^2 = \lambda^2 \left(J^\top J\right)_{\alpha\alpha} + (1 - \lambda^2) \left(J^\top J\right)_{\beta\beta} + 2\lambda\sqrt{1 - \lambda^2} \left(J^\top J\right)_{\alpha\beta} \tag{D.4}$$

To find, for given α and β, the sloppiest linear combination we can now take derivatives with respect to λ, set the function equal to zero, and solve for λ_*.

$$2\lambda_* \left(J^\top J\right)_{\alpha\alpha} - 2\lambda_* \left(J^\top J\right)_{\beta\beta} + \left(2\sqrt{1 - \lambda_*^2} - \frac{2\lambda_*^2}{\sqrt{1 - \lambda_*^2}}\right) \left(J^\top J\right)_{\alpha\beta} = 0. \tag{D.5}$$

This equation has four solutions related by taking $\lambda_* \to -\lambda_*$, $\lambda_* \to \sqrt{1 - \lambda_*^2}$, and $\lambda_* \to -\sqrt{1 - \lambda_*^2}$. Precisely which one yields the minimum is determined by the signs and magnitudes of the relevant components of $J^\top J$. Aside from these

transformations, the solution is

$$\lambda_* = -\sqrt{\frac{\sqrt{D_{\alpha\beta}} - (J^\top J)_{\alpha\alpha} + (J^\top J)_{\beta\beta}}{2\sqrt{D_{\alpha\beta}}}} \tag{D.6}$$

where $D_{\alpha\beta} = \left((J^\top J)_{\alpha\alpha} - (J^\top J)_{\beta\beta}\right)^2 + 4(J^\top J)^2_{\alpha\beta}$. For this linear combination of p_α and p_β, the squared sensitivity is given by

$$\left|\frac{\partial r_i}{\partial \tilde{p}}\right|^2 = \frac{-D_{\alpha\beta} + 4(J^\top J)^2_{\alpha\beta}}{2\sqrt{D_{\alpha\beta}}} + \\ \frac{1}{2}\left((J^\top J)_{\alpha\alpha} - 4(J^\top J)_{\alpha\beta}\sqrt{(J^\top J)^2_{\alpha\beta}/D_{\alpha\beta}} + (J^\top J)_{\beta\beta}\right). \tag{D.7}$$

We can now take the Hessians for some real-life models and look for particular linear combinations of pairs of parameters that cause substantial sloppiness. Both models we will consider are of biological networks. In both cases, derivatives were taken with respect to the logarithms of the biochemical reaction constants (rate and Michaelis-Menten constants). Because of these logarithms, the sum of any two parameters is equal to the product of the two biochemical constants and the difference of any two parameters is the ratio of the biochemical constants.

The first model we consider is for growth factor signaling network in PC12 (Section 1.2). In Figure D.1 we show the matrix of $|\partial r_i/\partial \tilde{p}|^2$ values for each pair of the 48 parameters. Figure D.2 shows the corresponding values of λ_*. While there is much information that could be gleaned from these plots, as a demonstration of the usefulness we will focus on the pair of parameters indexed by 42 and 18. Parameter 42 is the (logarithm of the) rate constant for activation of BRaf by Rap1 and parameter 18 is the (logarithm of the) rate constant for activation of Mek1/2 by BRaf. This analysis shows that the model is particularly insensitive to the ratio of these two rate constants ($\lambda_* = -0.709$ for the combination $\tilde{p} = \lambda p_{42} + \sqrt{1-\lambda^2}p_{18}$). This result confirms the findings in Appendix C that these two parameters lead to
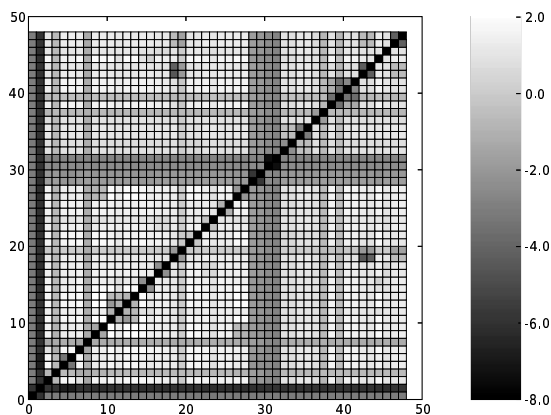
Figure D.1:   PC12 model sensitivity to parameter pairs. The horizontal and vertical axes are the indices of the 48 parameters in this model. The color is a log scale, $\log_{10}(|\partial r_i/\partial \tilde{p}|^2)$.

particularly sloppy submodels. This suggests that perhaps the network does not need both proteins to be in the model and could be simplified by lumping them into one effective mechanism for activating Mek. As further confirmation that this particular combination of the two rate constants constitutes a sloppy direction, in Figure D.3 we show the dot product of this direction with each of the eigenvectors of the Hessian (approximated by $J^\top J$). It is clear that while not precisely an eigenvector itself, this direction in parameter space falls well within the sloppy subspace and is not aligned with the stiff directions at all.

Since this new measure, $|\partial r_i/\partial \tilde{p}|^2$ defines how close any two parameters are to one another, we can use it to cluster the parameters. This can help us determine which sets of parameters consitute Vandermonde subsystems because they will all cluster together. In Figure D.4 we show the results of clustering the PC12 network parameters in just this way. On the top we see the original Jacobian, $J_{ij} = \partial r_i/\partial p_j$,
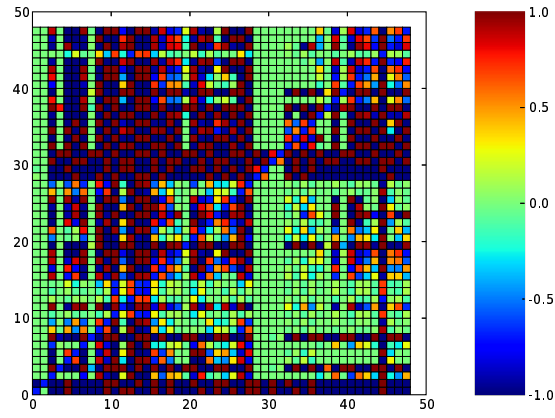
93

Figure D.2: Insensitive parameter pairs in PC12 model. The horizontal and vertical axes are the indices of the 48 parameters in this model. The color scale represents λ where the linear combination of the given pair of parameters to which the model is most insensitive is $\tilde{p} = \lambda p_\alpha + \sqrt{1 - \lambda^2} p_\beta$ (α and β are the column and row indices, respectively).
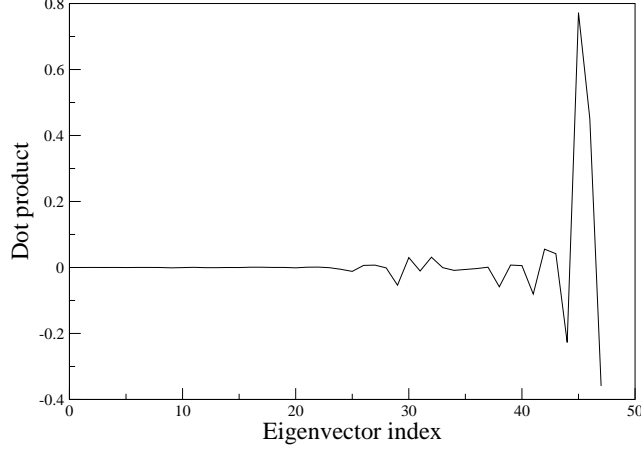
Figure D.3: PC12 sensitivity to $\tilde{p}$. The horizontal axis indexes the PC12 Hessian (approximated by $J^\top J$) eigenvectors with 0 being the stiffest and 48 being the sloppiest. The vertical scale is the dot product of $\tilde{p}$ with each eigenvector, where $\tilde{p} = \lambda p_\alpha + \sqrt{1 - \lambda^2} p_\beta$ ($\lambda = -0.709$, $\alpha = 42$, and $\beta = 18$).

and on the bottom we see the same Jacobian but with the columns permuted according to the results of clustering based on $|\partial r_i / \partial \tilde{p}|^2$. There is one important detail: clustering simply on $|\partial r_i / \partial \tilde{p}|^2$ would be dominated by single parameters that have no effect on the residuals by themselves. What we are interested in is identifying sets of parameters which have significant and redundant effects. For this reason divide by the effects of each parameter alone and cluster on:

$$
\begin{aligned}
\text{PairProximity}_{\alpha\beta} &= \frac{|\partial r_i / \partial \tilde{p}|^2}{\lambda^2 |\partial r_i / \partial p_\alpha|^2 + (1 - \lambda^2) |\partial r_i / \partial p_\beta|^2} \\
&= 1 + \frac{2\lambda \sqrt{1 - \lambda^2} \left(J^\top J\right)_{\alpha\beta}}{\lambda^2 \left(J^\top J\right)_{\alpha\alpha} + (1 - \lambda^2) \left(J^\top J\right)_{\beta\beta}}.
\end{aligned} \tag{D.8}
$$

In the hierarchical clustering process one must decide how to define the distance between any already clustered sets of parameters. Since we want to identify the closest neighbors in this space (as opposed to, for instance, centroids defining the clusters) we define the distance between two clusters as the minimum distance
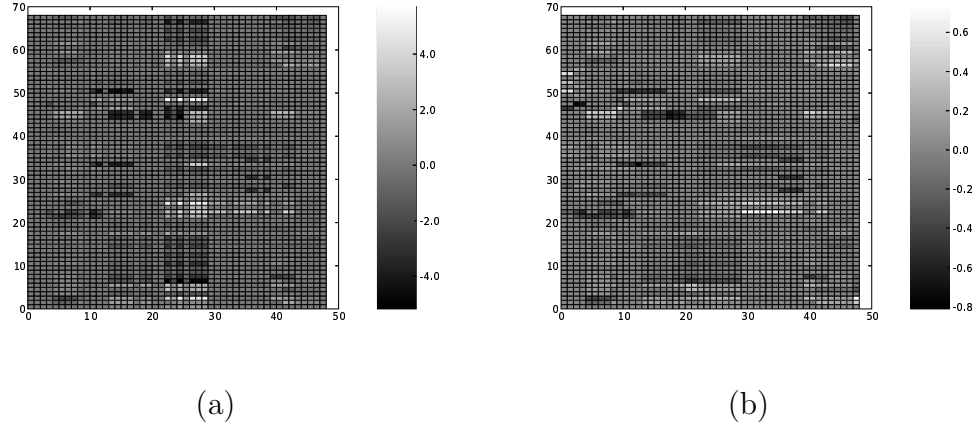
95

Figure D.4: Clustering the PC12 Jacobian parameters. In (a) the Jacobian matrix, $J_{ij} = \partial r_i / \partial p_j$ for the PC12 growth factor signaling network is shown. The columns have been ordered by a hierarchical clustering algorithm based on the 'distance' measure $|\partial r_i / \partial \tilde{p}|^2$. Note that there do in fact exist several sets of parameters which have equivalent patterns of effects on the model behavior and thus constitute Vandermonde subsystems. In (b) the same matrix is shown but the columns have each been normalized to have unit magnitude. This allows the effects of parameters which have little effect overall (such as the two left-most columns) to be noticeable to the eye.

between any two members of the clusters.

The next model that we analyze in this way is a model for signaling, trafficking and down-regulation based around the EGF receptor [6]. In Figure D.5 we plot $|\partial r_i / \partial \tilde{p}|^2$ values for the sloppiest combination of each pair of the 56 parameters. The particular combinations are depicted in Figure D.6, where we plot λ as a function of the parameter indices α and β for $\tilde{p} = \lambda p_\alpha + \sqrt{1 - \lambda^2} p_\beta$. The rows and columns which appear as dark stripes in Figure D.5 are individual parameters to which the model is insensitive. This is shown in Figure D.6 where the values of

λ_* are either -1, 0, or 1. These are also the very same parameters which lead to particularly sloppy submodels (Appendix C).

As with the previous model, we could analyze these matrices in much greater detail but for now we focus on a particular entry. Consider the pair of parameters indexed by 41 and 39. These are the (logarithms of the) Michaelis-Menten constants for activation of Focal Adhesion Kinase (FAK) by Src and for activation of Src by EGFR, respectively. Active FAK then activates Cbl, a particularly important protein in the model. Figures D.5 and D.6 show that the model is particularly insensitive to the ratio of these constants[1]. This is supported by Figure D.7 which plots the dot product of this direction in parameter space with each of the eigenvectors of the Hessian (approximated by $J^\top J$). These results suggest that the two proteins are not both necessary to recreate the experimental dynamics and that instead a single-step for activating Cbl by the EGFR would be sufficient.

In Figure D.8 we show the results of clustering the 56 parameters in this model. As before with the PC12 network, this clustering is based on the 'distance' defined by Equation D.8. Note that several sets of parameters are immediately identifiable as having similar patterns of effects on the residuals, suggesting that they constitute Vandermonde subsystems of the full model.

---

[1]Interestingly, the product of these two Michaelis-Menten constants appears in some of the stiffest eigenvectors [6].
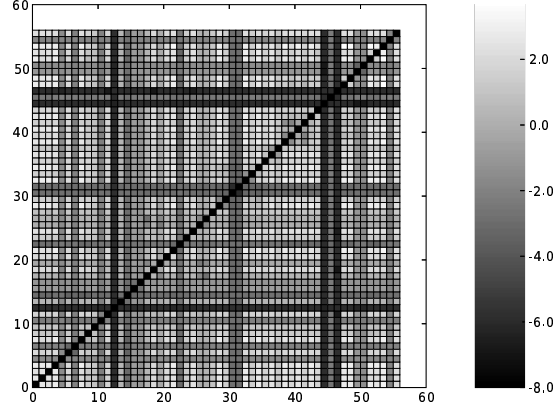
Figure D.5: EGFR model sensitivity to parameter pairs. The horizontal and vertical axes are the indices of the 56 parameters in this model. The color is a log scale, $\log_{10}(|\partial r_i/\partial \tilde{p}|^2)$.
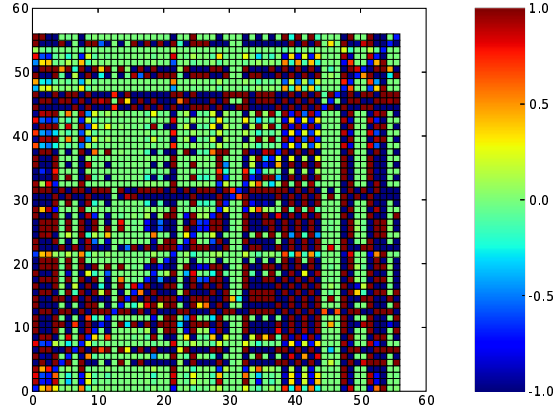

Figure D.6: Insensitive parameter pairs in EGFR model. The horizontal and vertical axes are the indices of the 56 parameters in this model. The color scale represents λ where the linear combination of the given pair of parameters to which the model is most insensitive is $\tilde{p} = \lambda p_\alpha + \sqrt{1-\lambda^2}p_\beta$ (α and β are the column and row indices, respectively).
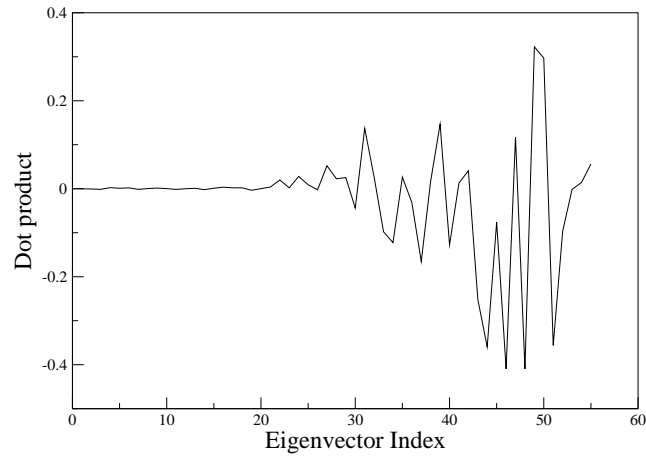
Figure D.7: EGFR sensitivity to $\tilde{p}$. The horizontal axis indexes the EGFR Hessian (approximated by $J^\top J$) eigenvectors with 0 being the stiffest and 56 being the sloppiest. The vertical scale is the dot product of $\tilde{p}$ with each eigenvector, where $\tilde{p} = \lambda p_\alpha + \sqrt{1 - \lambda^2} p_\beta$ ($\lambda = -0.834$, $\alpha = 41$, and $\beta = 39$).
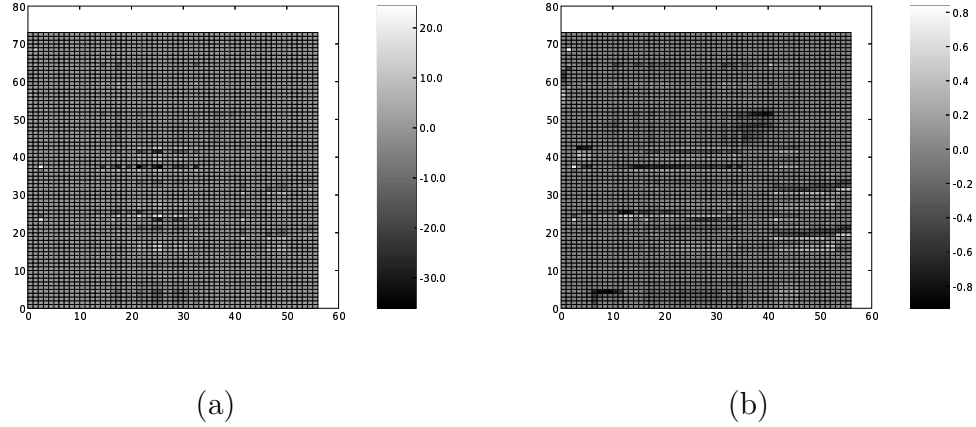
Figure D.8: Clustering the EGFR Jacobian parameters. In (a) the Jacobian matrix, $J_{ij} = \partial r_i / \partial p_j$ for the EGFR trafficking network is shown. The columns have been ordered by a hierarchical clustering algorithm based on the 'distance' measure $|\partial r_i / \partial \tilde{p}|^2$. Note that there do in fact exist several sets of parameters which have equivalent patterns of effects on the model behavior and thus constitute Vandermonde subsystems. In (b) the same matrix is shown but the columns have each been normalized to have unit magnitude. This allows the effects of parameters which have little effect overall (such as many of the right-most columns) to be noticeable to the eye.

Appendix E

Model Equations

The system of coupled ordinary first order differential equations that define the model of quorum sensing in *Agrobacterium tumefaciens* is as follows. Chemical species with names beginning with 'hot', as well as the variables 'radiolabel' and 'unradiolabel', are for radiolabeled pulse-chase experiments.

Differential Equations

$$\frac{d\,[\text{octopine}]}{dt} = 0$$

$$\frac{d\,[\text{OOHL}]}{dt} = \text{k}_{\text{OOHL}} \cdot [\text{TraI}]$$
$$- \text{k}_{\text{ROOHL}} \cdot [\text{FreeTraR}] \cdot [\text{OOHL}]$$
$$- \text{k}_{\text{dOOHL}} \cdot [\text{OOHL}]$$
$$- \text{k}_{\text{ROOHL}} \cdot [\text{hotFreeTraR}] \cdot [\text{OOHL}]$$

$$\frac{d\,[\text{traAPromoter}]}{dt} = \text{k}_{\text{doublingtime}} \cdot [\text{traAPromoter}]$$

$$\frac{d\,[\text{traAmRNA}]}{dt} = \text{k}_{\text{traAbasal}} \cdot [\text{traAPromoter}]$$

$$+ \frac{k_{tmA} \cdot [traAPromoter] \cdot [TraRDimer]}{([TraRDimer] + Km_{tmA})}$$

$$- k_{dmA} \cdot [traAmRNA]$$

$$\frac{d[TraA]}{dt} = k_{tpA} \cdot [traAmRNA]$$

$$- k_{dpA} \cdot [TraA]$$

$$\frac{d[traRPromoter]}{dt} = k_{doublingtime} \cdot [traRPromoter]$$

$$\frac{d[traRmRNA]}{dt} = k_{traRbasal} \cdot [traRPromoter]$$

$$+ \frac{k_{tmR} \cdot [traRPromoter] \cdot [octopine]}{([octopine] + Km_{tmR})}$$

$$+ k_{Plac} \cdot [Plac]$$

$$- k_{dmR} \cdot [traRmRNA]$$

$$\frac{d[FreeTraR]}{dt} = k_{tpR} \cdot [traRmRNA] \cdot [unradiolabel]$$

$$- k_{dpR} \cdot [FreeTraR]$$

$$- k_{ROOHL} \cdot [FreeTraR] \cdot [OOHL]$$

$$\frac{d[BoundTraR]}{dt} = k_{ROOHL} \cdot [FreeTraR] \cdot [OOHL]$$

$$- k_{dimR} \cdot [BoundTraR]^2 \cdot 2$$

$$\frac{d[TraRDimer]}{dt} = k_{dimR} \cdot [BoundTraR]^2$$

$$\frac{d[traIPromoter]}{dt} = k_{doublingtime} \cdot [traIPromoter]$$

$$\frac{d[traImRNA]}{dt} = k_{traIbasal} \cdot [traIPromoter]$$

$$+ \frac{k_{tmI} \cdot [traIPromoter] \cdot [TraRDimer]}{([TraRDimer] + Km_{tmI})}$$

$$- k_{dmI} \cdot [traImRNA]$$

$$\frac{d[TraI]}{dt} = k_{tpI} \cdot [traImRNA]$$

102

$$- \mathrm{k_{dpI}} \cdot [\mathrm{TraI}]$$

$$\frac{d\,[\mathrm{Plac}]}{dt} \quad = \quad 0$$

$$\frac{d\,[\mathrm{hotFreeTraR}]}{dt} \quad = \quad \mathrm{k_{tpR}} \cdot [\mathrm{traRmRNA}] \cdot [\mathrm{radiolabel}]$$
$$- \mathrm{k_{dpR}} \cdot [\mathrm{hotFreeTraR}]$$
$$- \mathrm{k_{ROOHL}} \cdot [\mathrm{hotFreeTraR}] \cdot [\mathrm{OOHL}]$$

$$\frac{d\,[\mathrm{hotBoundTraR}]}{dt} \quad = \quad \mathrm{k_{ROOHL}} \cdot [\mathrm{hotFreeTraR}] \cdot [\mathrm{OOHL}]$$
$$- \mathrm{k_{dimR}} \cdot [\mathrm{hotBoundTraR}]^2 \cdot 2$$

$$\frac{d\,[\mathrm{hotTraRDimer}]}{dt} \quad = \quad \mathrm{k_{dimR}} \cdot [\mathrm{hotBoundTraR}]^2$$

$$\frac{d\,[\mathrm{radiolabel}]}{dt} \quad = \quad 0$$

$$\frac{d\,[\mathrm{unradiolabel}]}{dt} \quad = \quad 0$$

Appendix F

Fits and Eigenvectors

Figures F.1,F.2,F.3,F.4,F.5, and F.6 contain the data used to constrain the model as well as the simulation output with the best fit parameters. Table F.1 contains the best fit parameter values. The eigenvectors of $J^\top J$, sorted by their corresponding eigenvalues, are available in Figures F.7, F.8 and F.9.

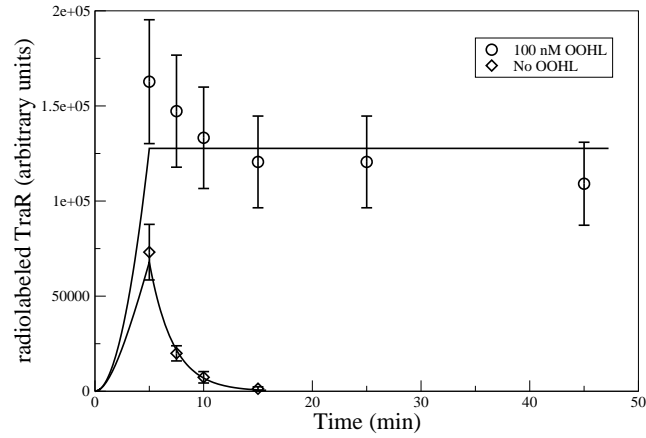Figure F.1: TraR protein half-life with and without OOHL [39]. Radiolabeled TraR translation was carried out in either the presence or absence of 100 nM OOHL. Without OOHL TraR is very unstable and has a half life of roughly two minutes. When OOHL is present during translation, TraR binds and is stable for the length of the experiment. The circles and error bars are the experimental measurements and the straight lines are the model output with the best fit parameters.
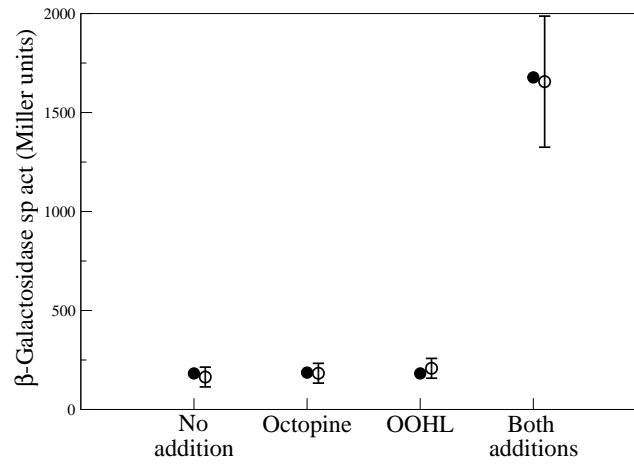
Figure F.2: Activation of quorum sensing requires both octopine and OOHL [11]. In wild type *Agrobacterium tumefaciens* the *lacZ* gene was put under TraR control. When present, octopine concentration was 2 mg/ml and OOHL concentration was 0.5 nM. The open circles and error bars are the experimental measurements and the solid circles are the model output with the best fit parameters.
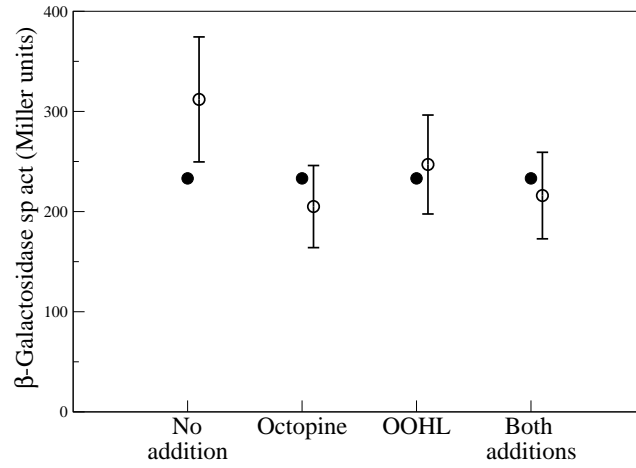
Figure F.3: Activation of quorum sensing requires the gene *traR* [11]. The *lacZ* gene was put under control of TraR in *Agrobacterium tumefaciens.* All experiments in this table were with a strain that had a disruption in the traR gene. When present, octopine concentration was 2 mg/ml and OOHL concentration was 0.5 nM. The open circles and error bars are the experimental measurements and the solid circles are the model output with the best fit parameters.
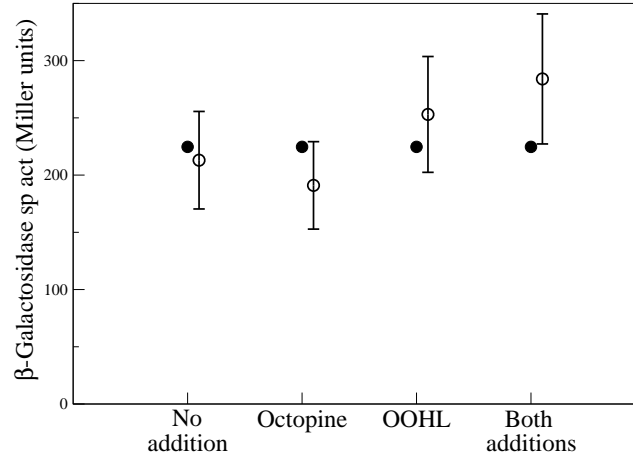
Figure F.4: Activation of quorum sensing requires the gene occR [11]. The *lacZ* gene was put under control of TraR in *Agrobacterium tumefaciens*. All experiments in this table were with a strain that had a disruption in the occR gene. When present, octopine concentration was 2 mg/ml and OOHL concentration was 0.5 nM. The open circles and error bars are the experimental measurements and the solid circles are the model output with the best fit parameters.
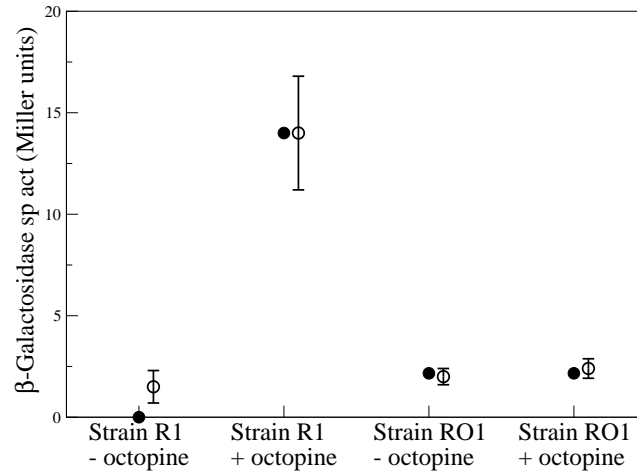
Figure F.5: Activation of traR expression requires the occR gene [11]. *Agrobacterium* strain R1 is wild-type and strain RO1 has a disruption in the occR gene. Read out of activation is a traR-lacZ fusion. The open circles and error bars are the experimental measurements and the solid circles are the model output with the best fit parameters.
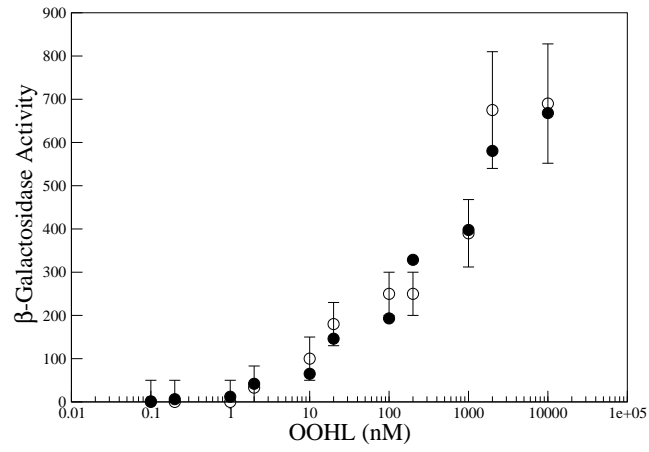
Figure F.6:   Dose response curve for activation of quorum sensing network in response to various concentrations of OOHL [38]. Readout for activation is lacZ gene under control of traA promoter. The open circles and error bars are the experimental measurements and the solid circles are the model output with the best fit parameters.

Table F.1: Best fit parameter values. Parameters beginning with a lower case 'k' have units of inverse minutes and 'Km' parameters have units of molecules per cell. Note that precisely because the system is sloppy, these values are not to be interpreted as the 'true' values. For instance, the basal transcription rate for traR of $4.31 \times 10^{-29} \text{min}^{-1}$ is equivalent to roughly 1 transcribed traR mRNA every ten millenia. The model with this set of training data simply needs this rate to be small (effectively zero) and the optimization algorithm has allowed it to evaporate to unbiological ranges. Similarly the value for the Michaelis-Menten constant involved in transcription of traA mRNA of 2.29×10^{12} simply means that the model does not need this relationship to saturate.

| Index	Parameter	Value	Index	Parameter	Value
0	$k_{traRbasal}$	4.31e-29	14	k_{OOHL}	3.43e-11
1	k_{tmR}	6.99e11	15	$k_{traAbasal}$	3.22e-08
3	k_{dmR}	6.71e-16	16	k_{tmA}	1.47e8
4	k_{tpR}	6.20e-11	18	k_{dmA}	9.77e-09
5	k_{dpR}	0.457	19	k_{tpA}	1.02e-05
6	k_{ROOHL}	4.10e-06	20	k_{dpA}	2.74e-23
7	k_{dimR}	4.44e-4	21	k_{dOOHL}	6.08e-2
8	$k_{traIbasal}$	6.61e-08	22	k_{Plac}	1.72e15
9	k_{tmI}	6.28e10	23	$k_{doublingtime}$	8.65e-4
11	k_{dmI}	3.61e-15	2	Km_{tmR}	1.05e-07
12	k_{tpI}	4.59e3	10	Km_{tmI}	6.00e8
13	k_{dpI}	1.59e-05	17	Km_{tmA}	2.29e12
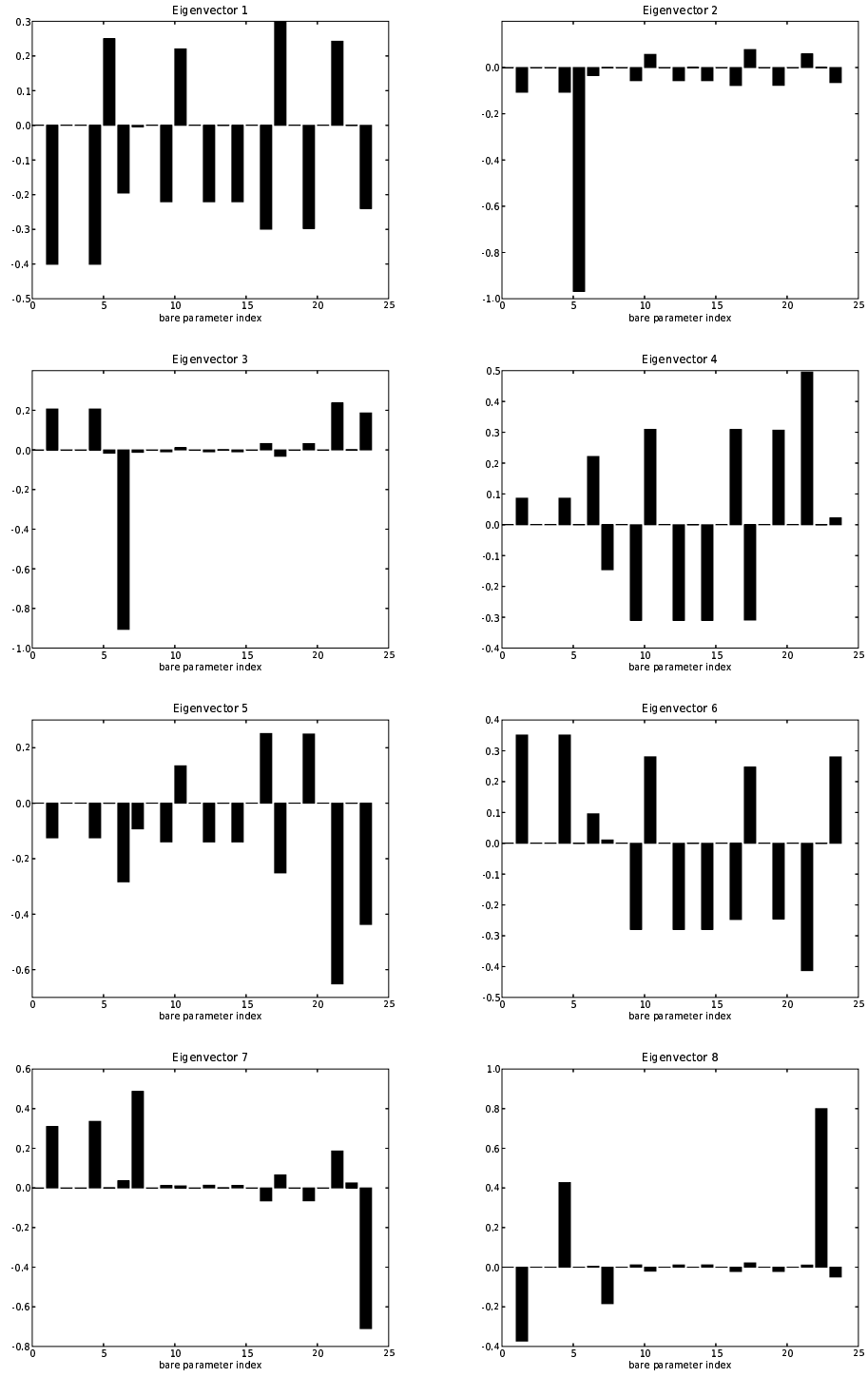
Figure F.7: First eight eigenvectors of quorum sensing Hessian. Eigenvectors one through eight of $J^\top J$ evaluated at the parameters in Table F.1. The eigenvectors are sorted by eigenvalue (Figure 2.4).
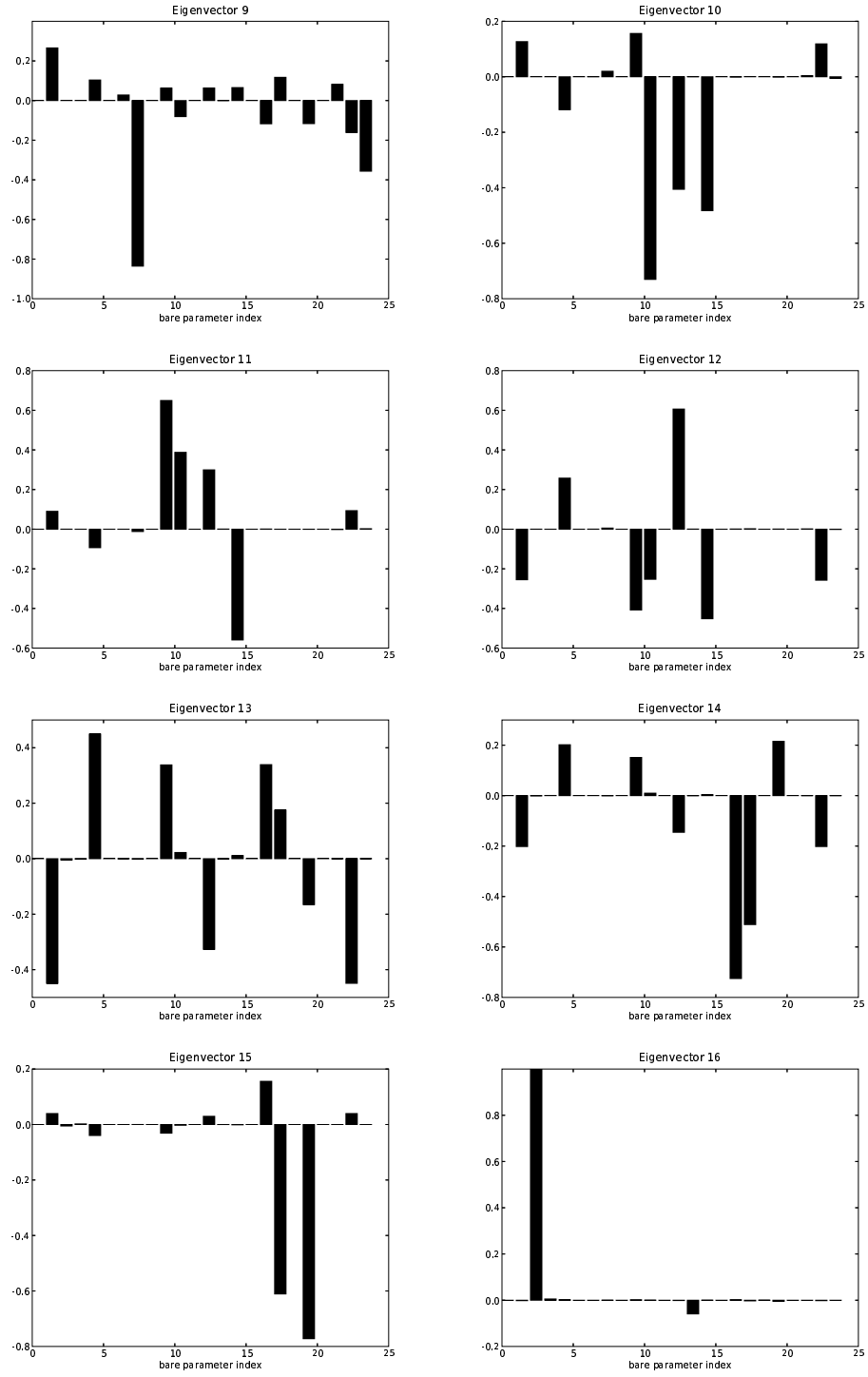
Figure F.8: Second eight eigenvectors of quorum sensing Hessian. Eigenvectors nine through sixteen of $J^\top J$ evaluated at the parameters in Table F.1. The eigenvectors are sorted by eigenvalue (Figure 2.4).
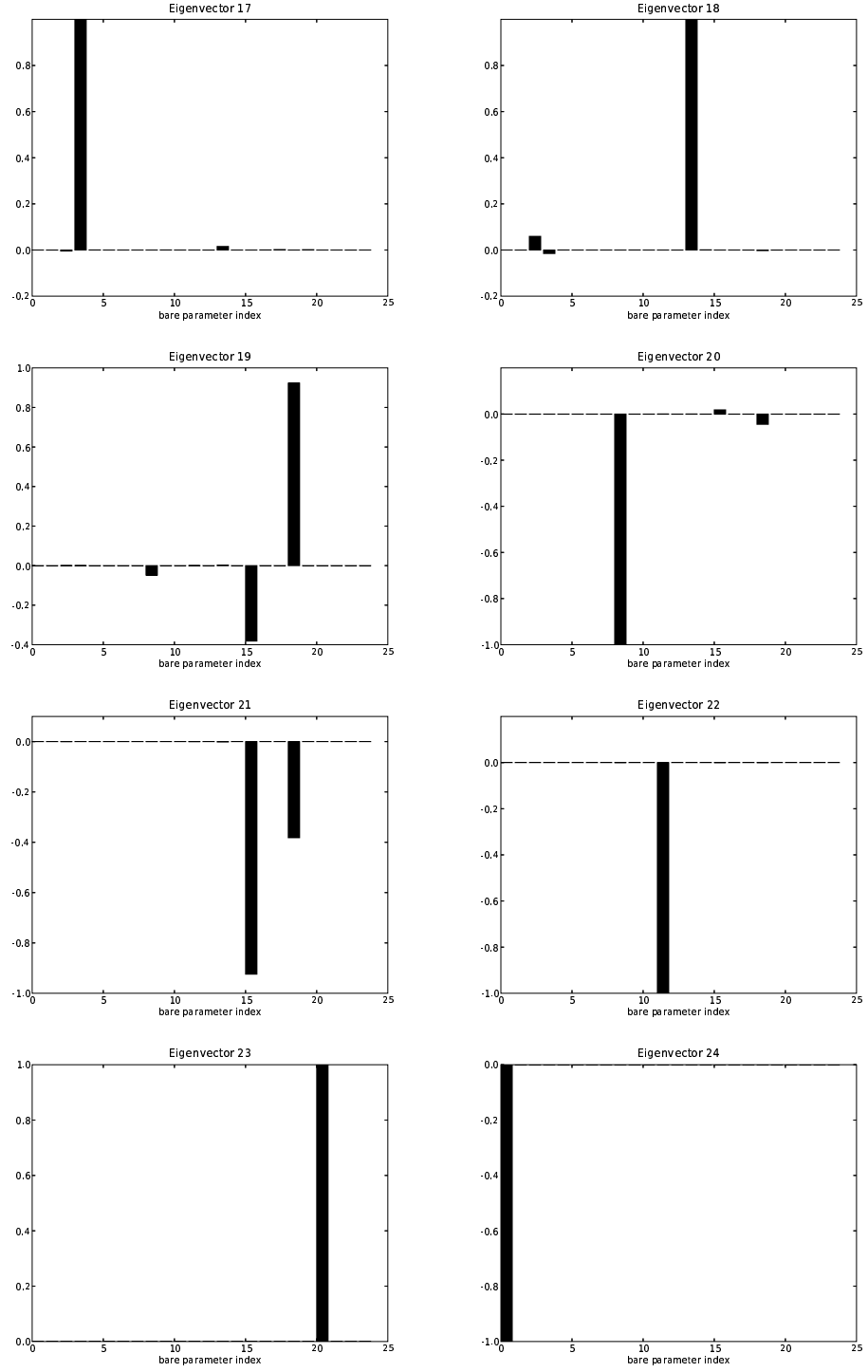
Figure F.9: Final eight eigenvectors of quorum sensing Hessian. Eigenvectors seventeen through twenty four of $J^\top J$ evaluated at the parameters in Table F.1. The eigenvectors are sorted by eigenvalue (Figure 2.4).

BIBLIOGRAPHY

[1] Reiko Akakura and Stephen C. Winans. Constitutive mutations of the occr regulatory protein affect dna bending in response to metabolites released from plant tumors. *J. Biol. Chem.*, 277(8):5866–5874, February 2002.

[2] James E. Bailey. Complex biology with no parameters. *Nat. Biotechnol.*, 19(6):503–504, June 2001.

[3] Carlo W.J. Beenakker. Random-Matrix Theory of Quantum Transport. *Rev. Mod. Phys.*, 69:731–815, 1997.

[4] Kevin S. Brown, Colin C. Hill, Guillermo A. Calero, Christopher R. Myers, Kelvin H. Lee, James P. Sethna, and Richard A. Cerione. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Phys. Biol.*, 1:184–95, 2004.

[5] Kevin S. Brown and James P. Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E.*, 68:021904, 2003.

[6] Fergal P. Casey, Daniel Baird, Qiyu Feng, Ryan N. Gutenkunst, Joshua J. Waterfall, Christopher R. Myers, Kevin S. Brown, Richard A. Cerione, and James P. Sethna. In preperation.

[7] Fergal P. Casey, Joshua J. Waterfall, Ryan N. Gutenkunst, Christopher R. Myers, and James P. Sethna. Submitted.

[8] Yunrong Chai, Jun Zhu, and Stephen C. Winans. Trlr, a defective trar-like protein of *agrobacterium tumefaciens*, blocks trar function in vitro by forming inactive trlr:trar dimers. *Mol. Microbiol.*, 40(2):414–421, April 2001.

[9] Katherine C Chen, Laurence Calzone, Attila Csikasz-Nagy, Frederick R Cross, Bela Novak, and John J Tyson. Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, 15(8):3841–3862, Aug 2004.

[10] Don B. Clewell, M. Victoria Franciab, Susan E. Flannaganb, and Florence Y. An. Enterococcal plasmid transfer: sex pheromones, transfer origins, relaxases, and the staphylococcus aureus issue. *Plasmid*, 48(3):193–201, Nov 2002.

[11] W. Claiborne Fuqua and Stephen C. Winans. A luxr-luxi type regulatory system activates *agrobacterium* ti plasmid conjugal transfer in the presence of a plant tumor metabolite. *J. Bacteriol.*, 176(10):2796–2806, May 1994.

[12] Clay Fuquae and Stephen C. Winans. Conserved cis-acting promoter elements are required for density-dependent transcription of *agrobacterium tumefaciens* conjugal transfer genes. *J. Bacteriol.*, 178(2):435–440, January 1996.

[13] Kapil G. Gadkar, Jeffrey Varner, and Francis J. Doyle III. Model identification of signal transduction networks from data using a state regulator problem. *IEEE Systems Biology*, 2:17–30, 2005.

[14] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3 edition, 1996.

[15] Ryan N. Gutenkunst. Personal communication.

[16] Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Submitted.

[17] David Hilbert. Ein beitrag zur theorie des Legendre'schen polynoms. *Acta mathematica*, 18:155–159, 1894.

[18] Kristen Kaasbjerg. Statistical Optimization of Quantum Mechanical Calculations: a Bayesian approach to error estimation in density-functional theory. Master's thesis, Technical University of Denmark, June 2005.

[19] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83(7):1467–1470, Aug 1999.

[20] Cornelius Lanczos. *Applied Analysis*. Prentice Hall, Inc., Englewood Cliffs, N. J. , 1956.

[21] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sbornik*, 1:457–483, 1967.

[22] Madan L. Mehta. *Random Matrices*. Academic Press, Boston, 1991.

[23] Jens J. Mortensen, Kristen Kaasbjerg, Sørin L. Frederiksen, Jens K. Nørskov, James P. Sethna, and Karsten W. Jacobsen. Bayesian Error Estimation in Density Functional Theory. *Phys. Rev. Letters*, 95:216401, 2005.

[24] Peter M. Nightengale and Cyrus R. Umrigar, editors. *Quantum Monte Carlo Methods in Physics and Chemistry*, chapter 5, pages 129–160. Kluwer Academic Publishers, Boston, 1999.

[25] Katherine M. Pappas, Christine L. Weingart, and Stephen C. Winans. Chemical communication in proteobacteria: biochemical and structural studies of signal synthases and receptors required for intercellular signalling. *Mol. Microbiol.*, 53:755–769, August 2004.

[26] Katherine M. Pappas and Stephen C. Winans. A luxr-type regulator from a*grobacterium tumefaciens* elevates ti plasmid copy number by activating transcription of plasmid replication genes. *Mol. Microbiol.*, 48(4):1059–1073, May 2003.

[27] *Radioactive Decay Tables*, (2004), http://las.perkinelmer.com/content/TotallyRAD/page

[28] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C.* Cambridge University Press, New York, second edition, 1996.

[29] Rosemary J. Redfield. Is quorum sensing a side effect of diffusion sensing? *Trends Microbiol.*, 10(8):365–370, August 2002.

[30] Sørin L. Frederiksen, Karsten W. Jacobsen, Kevin S. Brown, and James P. Sethna. Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials. *Phys. Rev. Letters*, 93:165501, 2004.

[31] Thomas Guhr, Axel Müller-Groeling and Hans A. Weidenmüller. Random–matrix theories in quantum physics: common concepts. *Phys. Rep.*, 299:189–425, 1998.

[32] Cyrus R. Umrigar. Personal communication.

[33] Adriaan van den Bos and Joost Herman Swarte. Resolvability of the parameters of multiexponentials and other sum models. *IEEE Trans. Signal Processing*, 41:313–322, 1993.

[34] Christopher M. Waters and Bonnie L. Bassler. Quorum sensing: Cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.*, 21:319–346, November 2005.

[35] Stephen C. Winans. Personal communication.

[36] John Wishart. The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, 20:32–52, 1928.

[37] Hai-Bao Zhang, Lian-Hui Wang, and Lian-Hui Zhang. Genetic control of quorum-sensing signal turnover in *agrobacterium tumefaciens*. *Proc. Natl. Acad. Sci. U.S.A.*, 99(7):4638–4643, April 2002.

[38] Jun Zhu, John W. Beaber, Margret I. Moré, Clay Fuqua, Anatol Eberhard, and Stephen C. Winans. Analogs of the autoinducer 3-oxooctanoyl-homoserine lactone strongly inhibit activity of the trar protein of *agrobacterium tumefaciens*. *J. Bacteriol.*, 180(20):5398–5405, October 1998.

[39] Jun Zhu and Stephen C. Winans. Autoinducer binding by the quorum-sensing regulator trar increases affinity for target promoters in vitro and decreases trar turnover rates in whole cells. *Proc. Natl. Acad. Sci. U.S.A.*, 96(9):4832–4837, April 1999.

[40] Jun Zhu and Stephen C. Winans. The quorum-sensing transcriptional regulator trar requires its cognate signaling ligand for protein folding, protease resistance, and dimerization. *Proc. Natl. Acad. Sci. U.S.A.*, 98(4):1507–1512, February 2001.