Physics 7653: Statistical Physics
http://www.physics.cornell.edu/sethna/teaching/653/
Material for Week 5
Exercises due Tuesday Sep 25
Last correction at September 18, 2018, 11:53 am

Pre-class Preparation

Thursday

Read: http://www.lassp.cornell.edu/sethna/pubPDF/SloppyReviewJCP.pdf, Sloppiness and Emergent Theories in Physics, Biology, and Beyond, Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna, *J. Chem. Phys.* **143**, 010901 (2015).

Answer: In section III, we see a hyperribbon giving the space of possible model predictions. In section IV, we argue that models on an edge or corner (or hyper-edge) of the model manifold correspond to simpler models – ones where a parameter or product of powers of parameters typically goes to zero or infinity. *Suppose the true system behavior lies in the interior of the model manifold, but our experimental measurement of the system has noise large compared to the thicknesses of the model manifold. Would the best fit model to our data usually be on an edge? What if the noise is large compared to the thinnest ten directions measured near the true model behavior?*

Submit electronically by 8:30 Thursday morning.

Tuesday

Browse: http://www.lassp.cornell.edu/sethna/Sloppy/ Sloppy Models Web pages.

Exercises

1. **Sloppy minimization.** (Statistics) ③

   "With four parameters I can fit an elephant. With five I can make it waggle it's trunk." This statement, attributed to many different sources (from Carl Friedrich Gauss to Fermi), reflects the problems found in fitting multiparameter models to data. One almost universal problem is *sloppiness* – the parameters in the model are poorly constrained by the data.

   Consider the classic ill-conditioned problem of fitting exponentials to radioactive decay data. If you know that at $t = 0$ there are equal quantities of N radioactive materials with half-lives θ_α, the radioactivity that you would measure is

$$y_{\boldsymbol{\theta}}(t) = \sum_{\alpha=0}^{N-1} \theta_\alpha \exp(-\theta_\alpha t). \tag{1}$$

Now, suppose you do not know the decay rates θ_α. Can you reconstruct them by fitting the data to experimental data $d(t)$?

Start with just two radioactive decay elements $N = 2$. Suppose the actual decay constants for $d(t)$ are $\boldsymbol{\theta}_0 = [1, 2]$ (so the data fall on the curve $d(t) = \exp(-t) + 2\exp(-2t) = y_{\boldsymbol{\theta}_0}(t)$). For convenience, suppose we have perfect data at all times, with uniform error bars, so the cost is an integral over all times of the square of the error

$$C[\boldsymbol{\theta}] = \int_0^\infty \left(y_{\boldsymbol{\theta}}(t) - d(t)\right)^2 \, \mathrm{d}t. \tag{2}$$

(a) *Draw a contour plot of C in the square $0.5 < \theta_\alpha < 2.5$, with contours at $C = \{2^{-12}, 2^{-11}, \ldots, 2^0\}$. You may need to set the number of grid points per side to 40 to see the two minima.*

One can see from the contour plot that measuring the two rate constants separately would be a challenge. This is because the two exponentials have similar shapes, so increasing one decay rate and decreasing the other can almost perfectly compensate for one another.

(b) *If we assume both elements decay with the same decay rate $\theta = \theta_0 = \theta_1$, minimize the cost to find the optimum choice for θ. Where is this point on the contour plot? Plot $d(t)$ and $y(t)$ with this single-exponent best fit on the same graph, over $0 < t < 2$. Do you agree that it would be difficult to distinguish these two fits?*

This problem can become much more severe in higher dimensions. The banana-shaped ellipses in your contour plot can become needle-like, with aspect ratios of more than a thousand to one (about the same as a human hair). The relative widths of the ellipses are given by the square roots of the eigenvalues of the cost C.

(c) *For our exercise, where the data are perfectly fit by $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, show that the cost Hessian is a continuous integral*

$$H_{\alpha\beta} = (J^T J)_{\alpha\beta} = J_{t\alpha} J_{t\beta} = \int_0^\infty J(t,\alpha) J(t,\beta) \mathrm{d}t \tag{3}$$

where the Jacobian is now the $\infty \times N$ 'matrix' $J(t, \alpha) = \exp(-\theta_\alpha t)(1 - \theta_\alpha t)$.

(d) *Write a routine to calculate $H(\boldsymbol{\theta})$ by doing the indefinite integral in eqn 3. Find the eigenvalues and the eigenvectors for the cost Hessian H for your plot in part (b), evaluated at $\boldsymbol{\theta}_0$, and check them against your contour plot. What is the ratio of the long axis to the short axis, as predicted from your eigenvalues? For a sum of nine exponentials, with $\boldsymbol{\theta}_0 = [1, 2, 3, \ldots, 9]$, construct the Hessian, find its eigenvalues. By what factor does each successive eigenvalue shrink? Are they sloppy (roughly equally spaced in log)?*

2. **Sloppy monomials.**[1] (Statistics) ③

The same function $f(x)$ can be approximated in many ways. Indeed, the same function can be fit in the same interval by the same type of function in several different ways! For example, in the interval $[0, 1]$, the function $\sin(2\pi x)$ can be approximated (badly) by a fifth-order Taylor expansion, a Chebyshev polynomial, or a least-squares (Legendre[2]) fit:

$$\begin{aligned} f(x) &= \sin(2\pi x) \\ f_{\text{Taylor}} &\approx 0.000 + 6.283x + 0.000x^2 - 41.342x^3 \\ &\quad + 0.000x^4 + 81.605x^5 \\ f_{\text{Chebyshev}} &\approx 0.007 + 5.652x + 9.701x^2 - 95.455x^3 \\ &\quad + 133.48x^4 - 53.39x^5 \\ f_{\text{Legendre}} &\approx 0.016 + 5.410x + 11.304x^2 - 99.637x^3 \\ &\quad + 138.15x^4 - 55.26x^5 \end{aligned}$$

It is not a surprise that the best fit polynomial differs from the Taylor expansion, since the latter is not a good approximation. But it is a surprise that the last two polynomials are so different. The maximum error for Legendre is less than 0.02, and for Chebyshev is less than 0.01, even though the two polynomials differ by

$$\begin{aligned} &\text{Chebyshev} - \text{Legendre} = \\ &\quad - 0.009 + 0.242x - 1.603x^2 \\ &\quad + 4.182x^3 - 4.67x^4 + 1.87x^5 \end{aligned} \tag{4}$$

a polynomial with coefficients two hundred times larger than the maximum difference!

This flexibility in the coefficients of the polynomial expansion is remarkable. We can study it by considering the dependence of the quality of the fit on the parameters. Least-squares (Legendre) fits minimize a cost C^{poly}, the integral of the squared difference between the polynomial and the function:

$$\begin{aligned} C^{\text{poly}} &= (1/2) \int_0^1 (f(x) - y_{\boldsymbol{\theta}}(x))^2 \, \mathrm{d}x, \\ y_{\boldsymbol{\theta}}(x) &= \sum_{m=0}^{M} \theta_m x^m \end{aligned} \tag{5}$$

How quickly does this cost increase as we move the parameters θ_m away from their best-fit values? Varying any one monomial coefficient will of course make the fit bad. But

[1] Thanks to Joshua Waterfall, whose research is described here.

[2] The orthogonal polynomials used for least-squares fits on [-1,1] are the Legendre polynomials, assuming continuous data points. Were we using orthogonal polynomials for this exercise, we would need to shift them for use in [0,1].

apparently certain coordinated changes of coefficients do not cost much – for example, the difference between least-squares and Chebyshev fits given in eqn 4.

How should we explore the dependence in arbitrary directions in parameter space? We can use the eigenvalues of the Hessian to see how sensitive the fit is to moves along the various eigenvectors. . .

(a) *Note that the first derivative of the cost* C^{poly} *is zero at the best fit. Show that the Hessian second derivative of the cost is*

$$H_{mn}^{\text{poly}} = \frac{\partial^2 C^{\text{poly}}}{\partial\theta_m \partial\theta_n} = \frac{1}{m+n+1}. \tag{6}$$

This Hessian is the Hilbert matrix, famous for being ill-conditioned (having a huge range of eigenvalues).[3] Tiny eigenvalues of H^{poly} correspond to directions in polynomial space where the fit does not change.

(b) *Calculate the eigenvalues of the* 6×6 *Hessian for fifth-degree polynomial fits. Do they indeed span a large range? How big is the condition number (the ratio of the largest to the smallest eigenvalue)? Calculate the eigenvalues of larger Hilbert matrices. At what size do your eigenvalues seem contaminated by rounding errors?*

Notice from Eqn 6 that the dependence of the polynomial fit on the monomial coefficients is *independent of the function* $f(x)$ being fitted. We can thus vividly illustrate the sloppiness of polynomial fits by considering fits to the *zero* function $f(x) \equiv 0$. A polynomial given by an eigenvector of the Hilbert matrix with small eigenvalue must stay close to zero everywhere in the range $[0, 1]$. Let us check this.

(c) *Calculate the eigenvector corresponding to the smallest eigenvalue of* H^{poly}*, checking to make sure its norm is one (so the coefficients are of order one). Plot the corresponding polynomial in the range* $[0, 1]$*: does it stay small everywhere in the interval? Plot it in a larger range* $[-1, 2]$ *to contrast its behavior inside and outside the fit interval.*

This turns out to be a fundamental property that is shared with many other multiparameter fitting problems. Many different terms are used to describe this property. The fits are called *ill-conditioned*: the parameters θ_n are not well constrained by the data. The *inverse problem* is challenging: one cannot practically extract the parameters from the behavior of the model. Or, as our group describes it, the fit is *sloppy*: only a few directions in parameter space (eigenvectors corresponding to the largest eigenvalues) are constrained by the data, and there is a huge space of models (polynomials) varying along sloppy directions that all serve well in describing the data.

At root, the problem with polynomial fits is that all monomials x^n have similar shapes on $[0, 1]$: they all start flat near zero and bend upward. Thus they can be traded for one another; the coefficient of x^4 can be lowered without changing the fit if the coefficients

[3] For fits to discrete data, at points $(x_1, \ldots, x_N)$, the Hessian $H^{\text{poly}} = V^T V$, where $V_{i\alpha} = x_i^\alpha$ is the Vandermonde matrix, also famous for being ill-conditioned.

of x^3 and x^5 are suitably adjusted to compensate. Indeed, if we change basis from the coefficients θ_n of the monomials x^n to the coefficients ℓ_n of the orthogonal (shifted Legendre) polynomials, the situation completely changes. The Legendre polynomials are designed to be different in shape (orthogonal), and hence cannot be traded for one another. Their coefficients ℓ_n are thus well determined by the data, and indeed the Hessian for the cost C^{poly} in terms of this new basis is the identity matrix.