Lecture 13: Many Cheerful Facts about the Laplacian

36-720, Fall 2016

17 October 2016

Abstract

I clean up some of the confusions I inadvertently created during lecture, and sketch a proof of a key fact.

Contents

1 Linear Dynamical Systems

Suppose we're dealing with the time-evolution of an n-dimensional vector $\vec{x} \in \mathbb{R}^n$. Saying that the evolution operator is **linear** is saying that there is some $n \times n$ matrix $\mathbf{c}$ such that

$$\vec{x}(t+1) = \mathbf{c}\vec{x}(t) \tag{1}$$

Given the initial condition $\vec{x}(0)$, the trajectory is uniquely determined by iterating Eq. 1.

Since $\mathbf{c}$ is an $n \times n$ matrix, it has at most n distinct eigenvalues[1] If there are n distinct eigenvalues, there are n orthogonal eigenvectors. If an eigenvalue

[1] Because the characteristic polynomial giving the eigenvalues is of order n, and hence has at most n distinct roots.

is d-fold degenerate, its eigenvectors form a linear subspace[2] of dimension at most d, and this subspace is orthogonal to all the other eigenspaces. Supposing (what is generically the case) that the dimension of this subspace is indeed d, we can choose d orthogonal vectors from that space to serve as "the" eigenvectors. We can also scale all these eigenvectors to have length 1. Thus, we can find n orthogonal, unit-length vectors, say $\vec{v}_1$ through $\vec{v}_n$, which are eigenvectors of $\mathbf{c}$, with eigenvalues λ_1 through λ_n (possibly with repetition). These eigenvectors form a basis for $\mathbf{R}^n$.

Since the eigenvectors form a basis, we can expand any vector in terms of them. In particular, we can re-write the initial condition $\vec{x}(0)$ as a weighted sum of eigenvectors:

$$\vec{x}(0) = \sum_{i=1}^{n} \vec{v}_i(\vec{v}_i \cdot \vec{x}(0)) \equiv \sum_{i=1}^{n} w_i \vec{v}_i \tag{2}$$

defining the weights in the last equation.

Now the dynamics are very simple. For one step,

$$\begin{aligned}
\vec{x}(1) &= \mathbf{c}\vec{x}(0) && (3)\\
&= \mathbf{c}\sum_{i=1}^{n} w_i \vec{v}_i && (4)\\
&= \sum_{i=1}^{n} w_i \mathbf{c}\vec{v}_i && (5)\\
&= \sum_{i=1}^{n} w_i \lambda_i \vec{v}_i && (6)\\
& && (7)
\end{aligned}$$

using the fact that linear operators commute with summation, and that $\vec{v}_i$ is an eigenvector. Iterating, over arbitrarily many steps,

$$\vec{x}(t) = \sum_{i=1}^{n} w_i \lambda_i^t \vec{v}_i \tag{8}$$

Notice that if $|\lambda_i| < 1$, then $\lambda_i^t \to 0$. That is, eigenvectors of $\mathbf{c}$ which correspond to eigenvalues within the unit circle represent directions in the state space which shrink exponentially fast. On the other hand, if $|\lambda_i| > 1$, then $|\lambda_i^t| \to \infty$ — the corresponding directions in the state space zoom off to infinity exponentially. Eigenvectors corresponds to eigenvalues on the unit circle are the only ones which stay constant in magnitude.

[2]Observe that if $\vec{u}$ and $\vec{v}$ are eigenvectors of $\mathbf{c}$ with the same eigenvalue λ, then $a\vec{u} + b\vec{v}$ is also an eigenvector with eigenvalue λ.

1.1 Related operators and their eigenvalues

The above discussion applies to the eigenvalues of the (linear) evolution operator in Eq. 1. It is of course equivalent to look at the *difference* between $\vec{x}(t+1)$ and $\vec{x}(t)$, the **increment** to the state:

$$\vec{x}(t+1) - \vec{x}(t) = \mathbf{c}\vec{x}(t) - \vec{x}(t) = (\mathbf{c} - \mathbf{I})\vec{x} \equiv \mathbf{b}\vec{x}(t) \tag{9}$$

The difference or increment operator $\mathbf{b}$ has the same eigenvectors as the evolution operator $\mathbf{c}$, but its eigenvalues are all offset by 1. To see this, consider what happens if $\vec{v}$ is an eigenvector of $\mathbf{c}$ with eigenvalue λ. Then

$$\mathbf{b}\vec{v} = (\mathbf{c} - \mathbf{I})\vec{v} = \lambda\vec{v} - \vec{v} = (\lambda - 1)\vec{v} \tag{10}$$

On the other hand, if $\vec{u}$ is an eigenvector of $\mathbf{b}$ with eigenvalue κ, then

$$\begin{aligned} \kappa\vec{u} &= (\mathbf{c} - \mathbf{I})\vec{u} && (11)\\ \kappa\vec{u} + \vec{u} &= \mathbf{c}\vec{v} && (12)\\ (\kappa + 1)\vec{u} &= \mathbf{c}\vec{u} && (13) \end{aligned}$$

In continuous time, instead of the difference or increment operator, we'd have a differential operator:

$$\frac{d\vec{x}}{dt} = \mathbf{b}\vec{x} \tag{14}$$

The solution is, of course,

$$\vec{x}(t) = e^{t\mathbf{b}}\vec{x}(0) \tag{15}$$

In this context, $\mathbf{b}$ is called the **generator** of the time-evolution.

Here the exponential of a matrix is to be understood through the power series:

$$e^{t\mathbf{b}} = \sum_{k=0}^{\infty} \frac{(t\mathbf{b})^k}{k!} \tag{16}$$

Notice that $\mathbf{b}^k$ will have the same eigenvectors as $\mathbf{b}$, and its eigenvalues will be the k^{th} powers of the eigenvalues of $\mathbf{b}$. If $\vec{v}$ is an eigenvector of $\mathbf{b}$ with eigenvalue λ, then

$$\begin{aligned} e^{t\mathbf{b}}\vec{v} &= \sum_{k=0}^{\infty} \frac{(t\mathbf{b})^k}{k!}\vec{v} && (17)\\ &= \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!}\vec{v} && (18)\\ &= e^{t\lambda}\vec{v} && (19) \end{aligned}$$

so $\vec{v}$ is also an eigenvector of $e^{t\mathbf{b}}$, with eigenvalue $e^{t\lambda}$.

1.2 Evolution of Probability Distributions for Markov Chains

In an n-state (homogeneous) Markov chain, we have an $n \times n$ **transition matrix** $\mathbf{p}$, where p_{ij} is the probability that, when the chain is in state i, it will next move to state j,

$$\mathbb{P}\left(X(t+1) = j | X(t) = i\right) = p_{ij} \tag{20}$$

The relationship between $X(t+1)$ and $X(t)$ can be very nonlinear[3]. However, there is always a linear system buried within this.

We can represent a distribution over the states by a vector $\vec{\rho} \in \mathbb{R}^n$, with the constraints that $\rho_i \geq 0$, $\sum_{i=1}^{n} \rho_i = 1$. Applying the Markov chain for one step will transform this distribution to a new one, as follows:

$$\begin{aligned}
\mathbb{P}\left(X(t+1) = i\right) &= \sum_{j=1}^{n} \mathbb{P}\left(X(t+1) = i, X(t) = j\right) & (21)\\
&= \sum_{j=1}^{n} \mathbb{P}\left(X(t+1) = i | X(t) = j\right) \mathbb{P}\left(X(t) = j\right) & (22)\\
&= \sum_{j=1}^{n} p_{ji} \rho_j(t) & (23)
\end{aligned}$$

We can imagine this as telling us about what will happen if we start a large number of independent, non-interacting copies of the Markov process, with the proportion of them begun in state i being ρ_i. In vector form, therefore,

$$\rho(t+1) = \rho(t)\mathbf{p} \tag{24}$$

We have seen equations like Eqs. 23 and 24 before, when looking at cumulative-advantage processes. They are often (especially in the physics literature) called **master equations**. Other names are **(Kolomogorov) forward** equations, or **Fokker-Planck** equations, though those terms, especially the latter, are sometimes reserved for the corresponding differential equations for continuous-time Markov processes.

Notice that the linear operator is now acting from the *right*, rather than the left as before. You can convince yourself that none of the reasoning above really depended on the linear operator acting from the left, so *distributions* over the state space will evolve linearly, with the components projecting on to eigenvectors within the unit circle shrinking exponentially. Since $\mathbf{p}$ is a **stochastic matrix**, i.e., one with non-negative entries where each row sums to 1, the Perron-Frobenius theorem tells us that its largest eigenvalue is 1, and the corresponding (left) eigenvector has non-negative entries. If the Markov chain is ergodic, then the eigenvalue 1 is non-degenerate; each ergodic component of the chain corresponds to a distinct eigenvector of eigenvalue 1.

[3] In expectation, or for some other measure of central tendency.

2 Averaging or Diffusion Dynamics on Networks

Suppose we have an undirected graph with adjacency matrix $\mathbf{a}$, and a field over the network, i.e., a vector $\vec{x} \in \mathbf{R}^n$. Let's suppose that the field evolves according to a very basic sort of averaging or diffusion dynamic, where if node i is next to node j and has a higher value of the field, it will re-distribute some of the field to j. A simple model of this form[4] is

$$x_i(t+1) - x_i(t) = r \sum_{j=1}^{n} a_{ij}(x_j(t) - x_i(t)) \tag{25}$$

There are two things worth noticing about this:

1. This is a linear dynamical system; if we multiply all the x values by any amount, the increments will multiply by the same factor.
2. The dynamics conserve the x-quantity. To see this, sum up all the changes to all nodes:

$$\begin{aligned} \sum_{i=1}^{n} x_i(t+1) - x_i(t) &= \sum_{i=1}^{n} r \sum_{j=1}^{n} a_{ij}(x_j(t) - x_i(t)) & (26) \\ &= r \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(x_j(t) - x_i(t)) & (27) \end{aligned}$$

   Since, in an undirected graph, $a_{ij} = a_{ji}$, this last sum must always be exactly zero. But the sum of increments is the increment to the sum, so the total amount of whatever x measures is left alone.

The conservation of x-stuff is one reason this kind of dynamic is called "diffusion"; think of a finite supply of some dye or perfume spreading out from node to node.

2.1 The Graph Laplacian

We can write Eq. 25 in vector form; since we've seen the dynamics are linear, for some $\mathbf{b}$,

$$\vec{x}(t+1) - \vec{x}(t) = \mathbf{b}\vec{x}(t) \tag{28}$$

Since, for each coordinate, the increment is proportional to r, we can say

$$\vec{x}(t+1) - \vec{x}(t) = -r\mathbf{L}\vec{x}(t) \tag{29}$$

for some $\mathbf{L}$.

Claim: $\mathbf{L} = \mathbf{d} - \mathbf{a}$, where $\mathbf{d}$ is the diagonal matrix of degrees of nodes. *Proof:* For node i, there are $\sum_j a_{ij} = d_i$ non-zero terms in the sum in Eq. 25. Each of them picks up a negative $x_i(t)$ term; the minus sign in 29 flips this around, so we need $\mathbf{d} - \mathbf{a}$.

The matrix $\mathbf{L}$ is the **graph Laplacian**.

[4] But not the only possible one, even for linear dynamics; see Exercise 2.

2.1.1 Connection to the Laplacian Differential Operator

In 3-D calculus, the Laplacian is a second-order differential operator,

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{30}$$

with counterparts in any number of dimensions. To see how this connects to the graph Laplacian, let's drop down to one dimension, where the Laplacian operator is just the second derivative, d^2/dx^2. Since differentiation is the limit of difference slopes,

$$\frac{df}{dx}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{31}$$

the second derivative is a limit of second differences

$$\frac{d^2 f}{dx^2}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x) - [f(x) - f(x-h)]}{h^2} = \lim \frac{[f(x+h) - f(x)] + [f(x-h) - f(x)]}{h^2} \tag{32}$$

Clearly, the n-dimensional Laplacian will work the same way, only with more notation than I feel like writing out at this point.

If our make our graph by setting down a grid of points in $\mathbf{R}^n$ with a separation of h between points, the graph Laplacian will give us the "second difference" in the numerator of Eq. 32. It is thus intuitively clear that, for such "geometric" graphs, the Laplacian differential operator is some sort of (scaled) limit of the graph Laplacians as $h \to 0$; explaining how matrices can have as their limit a continuous operator is more involved than I feel like going into right now, but it *can* be done[5], and this can be made rigorous. The argument doesn't just work for finite-dimensional Euclidean spaces. On other manifolds, the equivalent of the Laplacian is the **Laplace-Beltrami operator**, and the Laplacians of geometric graphs derived from those manifolds approach these differential operators as the grid size shrinks.

2.1.2 Spectral Properties of the Graph Laplacian

The spectrum — the eigenvalues and eigenvectors — of the Laplacian encodes an awful lot of information about the graph[6]. Here are some very basic facts:

1. Since $\mathbf{L}$ is symmetric, the eigenvalues of $\mathbf{L}$ are all real, i.e., not complex. Consequently, they can be unambiguously ordered, from the smallest λ_1 to the largest λ_n.

2. The all-ones vector, $\vec{1}$, is always an eigenvector of $\mathbf{L}$, with eigenvalue 0. To see this, notice that if $x_i = 1$ for all i, then Eq. 25 is zero (no matter what r is). Hence $\mathbf{L}\vec{1} = 0$, no matter what the graph.

[5] See, e.g., Ethier and Kurtz (1986).

[6] Chung (1997) is a good introduction, though some sections presume a reader who knows more differential geometry than is typical of statisticians.

3. If C is a connected component of the graph, write $\vec{1}_C$ for the vector which is 1 on C and 0 elsewhere. It is easy to see that $\mathbf{L}\vec{1}_C = 0$, so each connected component of the graph has a corresponding 0 eigenvector.

4. If $\vec{v}$ is a 0 eigenvector of $\mathbf{L}$, then it is either proportional to one of the $\vec{1}_C$, or is a linear combination of those vectors. Thus there is a one-to-one correspondence between connected components and 0 eigenvectors.

5. All of the eigenvectors of $\mathbf{L}$ are positive[7]. This is mostly easily seen by verifying (EXERCISE 1) that for any vector $\vec{x}$

$$\vec{x}^T \mathbf{L} \vec{x} = \frac{1}{2} \sum_{i,j} a_{ij} (x_i - x_j)^2 \tag{33}$$

Since $\vec{1}$ is a 0-eigenvector of $\mathbf{L}$, it is also a eigenvector of $\mathbf{I} - r\mathbf{L}$ with eigenvalue 1. The part of the initial condition which projects onto $\vec{1}$ is therefore invariant over time. For sufficiently small r, all the other eigenvectors of $\mathbf{I} - r\mathbf{L}$ will be within the unit circle, and so the projections of the initial conditions on to the other eigenvectors will shrink exponentially fast.

In a graph with one connected component, the second-largest eigenvalue of $\mathbf{L}$, λ_2, goes along with a vector, $\vec{v}_2$. Since this is orthogonal to the zero eigenvector $\vec{1}$, we must have $\vec{v}_2 \cdot \vec{1} = 0$, which in turn implies that $\vec{v}_2$ must contain entries of alternating signs. At large times t, then, we'll have

$$\vec{x}(t) \approx (\vec{x}(0) \cdot \vec{1})\vec{1} + (1 - r\lambda_2)^t (\vec{x}(0) \cdot \vec{v}_2)\vec{v}_2 \tag{34}$$

since all the terms along other eigenvectors are exponentially smaller. Thus any initial pattern will converge towards uniformity, and $\vec{v}_2$ tells us abut the most persistent non-uniformity — the split between nodes with positive and negative entries in $\vec{v}_2$ is the split which takes the longest time to average away.

2.2 Laplacian-based Smoothing

Suppose we observe a field x over a graph, and we think this represents some underlying true signal plus noise. If we suppose the signal is smooth over the graph, i.e., that nearby points have similar values, then we might estimate the signal by solving the following optimization problem:

$$\min_{\vec{f} \in \mathbb{R}^n} \|\vec{f} - \vec{x}\|^2 + \gamma \sum_{i,j} a_{ij} (f_j - f_i)^2 \tag{35}$$

By the magic of Lagrange multipliers, this is equivalent to minimizing the sum of squared errors $\|\vec{f} - \vec{x}\|^2$ under a constraint of the form $\sum_{i,j} a_{ij}(f_j - f_i)^2 \leq c$; the penalty factor γ is the Lagrange multiplier enforcing the constraint, i.e., the "shadow price" (in units of squared error) we'd pay to loosen the constraint.

[7] I did indeed screw this up in lecture.

One can show (Exercise 1) that

$$\sum_{i,j} a_{ij}(f_j - f_i)^2 = 2\vec{f}^T \mathbf{L} \vec{f} \tag{36}$$

Thus the optimal vector is

$$\begin{aligned} \hat{x} &= \operatorname*{argmin}_{\vec{f} \in \mathbb{R}^n} \|\vec{f} - \vec{x}\|^2 + 2\gamma \vec{f}^T \mathbf{L} \vec{f} && (37) \\ &= (I + 2\gamma \mathbf{L})^{-1} \vec{x} && (38) \end{aligned}$$

This resembles both ridge regression and spline smoothing, especially the latter, since $\mathbf{L}$ is so close to the second derivatives used to define smoothing splines. See Wehbe *et al.* (2015) for an application of this idea; Li *et al.* (2016) for an extension to allow for node-level covariates; and Corona *et al.* (2008) for a proper extension of splines on graphs.

3 Exercises

1. *Positive-definiteness of the graph Laplacian* To show that the graph Laplacian is positive-definite, i.e., that

   $$\vec{v}^T \mathbf{L} \vec{v} \geq 0 \tag{39}$$

   for any vector $\vec{v}$, it will be enough to show that this quadratic form is equal to a weighted sum of squares:

   $$\vec{v}^T \mathbf{L} \vec{v} = 2 \sum_{i,j} a_{ij}(v_j - v_i)^2 \tag{40}$$

   There are actually many ways to do this; here is one.

   (a) Show that

   $$\sum_{i,j} a_{ij}(v_j - v_i)^2 = \sum_i v_i^2 \sum_j a_{ij} + \sum_j v_j^2 \sum_i a_{ij} - 2 \sum_{i,j} v_i v_j a_{ij} \tag{41}$$

   (b) Show that

   $$\sum_i v_i^2 \sum_j a_{ij} = \sum_j v_j^2 \sum_i a_{ij} = \vec{v}^T \mathbf{d} \vec{v} \tag{42}$$

   (c) Show that

   $$\sum_{i,j} a_{ij}(v_j - v_i)^2 = 2\vec{v}^T (\mathbf{d} - \mathbf{a}) \vec{v} \tag{43}$$

2. For each node i, write $\overline{x}_{\mathcal{N}(i)}$ for the mean value of i's neighbors, i.e., the mean of the x_j where $a_{ij} = 1$.

(a) Show that Eq. 25 is equivalent to

$$x_i(t+1) - x_i(t) = r d_i(\overline{x}_{\mathcal{N}(i)}(t) - x_i(t)) \tag{44}$$

Notice that this means that if two nodes are equally far from the average of their neighbors, the higher-degree node will change more quickly.

(b) Consider instead the equation

$$x_i(t+1) - x_i(t) = r(\overline{x}_{\mathcal{N}(i)}(t) - x_i(t)) \tag{45}$$

Write this in matrix form, using the Laplacian and the $\mathbf{d}$ matrix. Call this increment operator $\mathbf{b}$.

(c) Explain how the eigenvalues and eigenvectors of $\mathbf{b}$ are related to those of the Laplacian.

References

Chung, Fan R. K. (1997). *Spectral Graph Theory*. Providence, Rhode Island: American Mathematical Society. URL `http://www.math.ucsd.edu/~fan/research/revised.html`.

Corona, Eduardo, Terran Lane, Curtis Storlie and Joshua Neil (2008). *Using Laplacian Methods, RKHS Smoothing Splines and Bayesian Estimation as a framework for Regression on Graph and Graph Related Domains*. Tech. Rep. TR-CS-2008-06, Department of Computer Science, University of New Mexico. URL `http://www.cs.unm.edu/~treport/tr/08-06/laplacian-rkhs.pdf`.

Ethier, Stewart N. and Thomas G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. New York: Wiley.

Li, Tianxi, Elizaveta Levina and Ji Zhu (2016). "Prediction models for network-linked data." arxiv:1602.01192. URL `https://arxiv.org/abs/1602.01192`.

Wehbe, Leila, Aaditya Ramdas, Rebecca C. Steorts and Cosma Rohilla Shalizi (2015). "Regularized Brain Reading with Shrinkage and Smoothing." *Annals of Applied Statistics*, **9**: 1997–2022. URL `http://arxiv.org/abs/1401.6595`. doi:10.1214/15-AOAS837.