update April 11

III.2 Interlacing Eigenvalues and Low Rank Signals

The previous section found the change in A^{-1} produced by a change in A. We could allow infinitesimal changes dA and also finite changes $\Delta A = -UV^{\mathrm{T}}$. The results were an infinitesimal change or a finite change in the inverse matrix :

$$\frac{dA^{-1}}{dt} = -A^{-1}\frac{dA}{dt}A^{-1} \quad \text{and} \quad \Delta A^{-1} = A^{-1}U(I - V^{\mathrm{T}}A^{-1}U)^{-1}V^{\mathrm{T}}A^{-1} \tag{1}$$

This section asks the same questions about the eigenvalues and singular values of A.

How do each λ and each σ change as the matrix A changes ?

You will see nice formulas for $d\lambda/dt$ and $d\sigma/dt$. But not much is linear about eigenvalues or singular values. Calculus succeeds for infinitesimal changes $d\lambda$ and $d\sigma$, because the derivative is a linear operator. But we can't expect to know exact values in the jumps to $\lambda(A + \Delta A)$ or $\sigma(A + \Delta A)$. Eigenvalues are more complicated than inverses.

Still there is good news. What can be achieved is remarkable. Here is a taste for a symmetric matrix S. Suppose S changes to $S + uu^{\mathrm{T}}$ (a "positive" change of rank 1). Its eigenvalues change from $\lambda_1 \geq \lambda_2 \geq \ \ldots$ to $z_1 \geq z_2 \geq \ \ldots$ We expect increases in eigenvalues since uu^{T} was positive semidefinite. *But how large are the increases ?*

Each eigenvalue z_i of $S + uu^{\mathrm{T}}$ is not smaller than λ_i or greater than λ_{i-1}. So the λ's and z's are "interlaced". Each $z_2, \ldots, z_n$ is between two λ's:

$$z_1 \geq \lambda_1 \geq z_2 \geq \lambda_2 \geq \ \ldots \ \geq z_n \geq \lambda_n. \tag{2}$$

We have upper bounds on the eigenvalue changes even if we don't have formulas for $\Delta\lambda$. There is one point to notice because it could be misunderstood. Suppose the change uu^{T} in the matrix is $Cq_2q_2^{\mathrm{T}}$ (where q_2 is the second unit eigenvector of S). Then $Sq_2 = \lambda_2 q_2$ will see a jump in that eigenvalue to $\lambda_2 + C$, because $(S + Cq_2q_2^{\mathrm{T}})q_2 = (\lambda_2 + C)q_2$. That jump is large if C is large. *So how could the second eigenvalue of $S + uu^{\mathrm{T}}$ possibly have $z_2 = \lambda_2 + C \leq \lambda_1$?*

Answer: If C is a big number, then $\lambda_2 + C$ is not the second eigenvalue of $S + uu^{\mathrm{T}}$! It becomes z_1, the **largest eigenvalue** of the new matrix $S + Cq_2q_2^{\mathrm{T}}$ (and its eigenvector is q_2). The original top eigenvalue λ_1 of S is now the second eigenvalue z_2 of the new matrix. So the statement (2) that $z_2 \leq \lambda_1 \leq z_1$ is the completely true statement (in this example) that $z_2 = \lambda_1$ is below $z_1 = \lambda_2 + C$.

We will connect this interlacing to the fact that the eigenvectors between $\lambda_1 = \lambda_{\max}$ and $\lambda_n = \lambda_{\min}$ are all **saddle points of the ratio $R(x) = x^{\mathrm{T}}Sx/x^{\mathrm{T}}x$**.

The Derivative of an Eigenvalue

We have a matrix $A(t)$ that is changing with the time t. So its eigenvalues $\lambda(t)$ are also changing. We will suppose that *no eigenvalues of* $A(0)$ *are repeated*—each eigenvalue $\lambda(0)$ of $A(0)$ can be safely followed for at least a short time t, as $\lambda(0)$ changes to an eigenvalue $\lambda(t)$ of $A(t)$. **What is its derivative $d\lambda/dt$?**

The key to $d\lambda/dt$ is to assemble the facts we know. The first is $A(t)\boldsymbol{x}(t) = \lambda(t)\boldsymbol{x}(t)$. The second is that the transpose matrix $A^{\mathrm{T}}(t)$ also has the eigenvalue $\lambda(t)$, because $\det(A^{\mathrm{T}} - \lambda I) = \det(A - \lambda I)$. Probably A^{T} has a different eigenvector $\boldsymbol{y}(t)$. When $\boldsymbol{x}$ is column k of the eigenvector matrix X for A, $\boldsymbol{y}$ is column k of the eigenvector matrix $(X^{-1})^{\mathrm{T}}$ for A^{T}, since $A = X^{-1}\Lambda X$ leads to $A^{\mathrm{T}} = X^{\mathrm{T}}\Lambda(X^{-1})^{\mathrm{T}}$. The lengths of $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalized by $X^{-1}X = I$. *This requires* $\boldsymbol{y}^{\mathrm{T}}(t)\boldsymbol{x}(t) = 1$ *for all* t.

When A is symmetric, its eigenvector $\boldsymbol{q}_1$ is the same for A and A^{T}.

Here are these facts on one line with the desired formula for $d\lambda/dt$ on the next line.

$$\boxed{\begin{array}{lccc} \textbf{Facts} & A(t)\,\boldsymbol{x}(t) = \boldsymbol{\lambda}(t)\,\boldsymbol{x}(t) & \boldsymbol{y}^{\mathrm{T}}(t)\boldsymbol{A}(t) = \boldsymbol{\lambda}(t)\boldsymbol{y}^{\mathrm{T}}(t) & \boldsymbol{y}^{\mathrm{T}}(t)\,\boldsymbol{x}(t) = 1 \\ \textbf{Formulas} & \boldsymbol{\lambda}(t) = \boldsymbol{y}^{\mathrm{T}}(t)\boldsymbol{A}(t)\,\boldsymbol{x}(t) & \text{and} & \dfrac{\boldsymbol{d\lambda}}{\boldsymbol{dt}} = \boldsymbol{y}^{\mathrm{T}}(t)\dfrac{\boldsymbol{dA}}{\boldsymbol{dt}}\,\boldsymbol{x}(t) \end{array}} \qquad \begin{array}{c}(3)\\(4)\end{array}$$

To find that formula $\lambda = \boldsymbol{y}^{\mathrm{T}}A\boldsymbol{x}$, just multiply the first fact $A\boldsymbol{x} = \lambda\boldsymbol{x}$ on the left side by $\boldsymbol{y}^{\mathrm{T}}$ and use $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} = 1$. Or multiply the second fact $\boldsymbol{y}^{\mathrm{T}}A = \lambda\boldsymbol{y}^{\mathrm{T}}$ on the right side by $\boldsymbol{x}$.

Now take the derivative of $\boldsymbol{\lambda} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\,\boldsymbol{x}$. The product rule gives three terms in $d\lambda/dt$:

$$\frac{d\lambda}{dt} = \frac{d\boldsymbol{y}^{\mathrm{T}}}{dt}A\,\boldsymbol{x} + \boxed{\boldsymbol{y}^{\mathrm{T}}\,\frac{\boldsymbol{dA}}{\boldsymbol{dt}}\,\boldsymbol{x}} + \boldsymbol{y}^{\mathrm{T}}A\,\frac{d\boldsymbol{x}}{dt} \tag{5}$$

The middle term is the correct derivative $d\lambda/dt$. The first and third terms add to zero:

$$\frac{d\boldsymbol{y}^{\mathrm{T}}}{dt}A\,\boldsymbol{x} + \boldsymbol{y}^{\mathrm{T}}A\,\frac{d\boldsymbol{x}}{dt} = \lambda\left(\frac{d\boldsymbol{y}^{\mathrm{T}}}{dt}\,\boldsymbol{x} + \boldsymbol{y}^{\mathrm{T}}\,\frac{d\boldsymbol{x}}{dt}\right) = \lambda\,\frac{d}{dt}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}\right) = \lambda\,\frac{d}{dt}(1) = 0. \tag{6}$$

Example If $A = \begin{bmatrix} 2t & 1 \\ 2t & 2 \end{bmatrix}$ then $\lambda^2 - 2(1+t)\lambda + 2t = 0$ and $\boldsymbol{\lambda} = \mathbf{1} + \boldsymbol{t} \pm \sqrt{\mathbf{1} + \boldsymbol{t}^2}$.

Then $\lambda_1 = \mathbf{2}$ and $\lambda_2 = \mathbf{0}$ at $t = 0$. The derivatives of λ_1 and λ_2 are $1 \pm t(1+t^2)^{-1/2} = \mathbf{1}$.

The eigenvectors for $\lambda_1 = 2$ at $t = 0$ are $\boldsymbol{y}_1^{\mathrm{T}} = \begin{bmatrix} 0 & 1 \end{bmatrix}$ and $\boldsymbol{x}_1 = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}$.

The eigenvectors for $\lambda_2 = 0$ at $t = 0$ are $\boldsymbol{y}_2^{\mathrm{T}} = \begin{bmatrix} 1 & -\frac{1}{2} \end{bmatrix}$ and $\boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Now equation (4) confirms that $\dfrac{\boldsymbol{d\lambda_1}}{\boldsymbol{dt}} = \boldsymbol{y}_1^{\mathrm{T}}\dfrac{\boldsymbol{dA}}{\boldsymbol{dt}}\boldsymbol{x}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} 1/2 \\ 1 \end{bmatrix} = 1.$

The Derivative of a Singular Value

A similar formula for $d\sigma/dt$ (derivative of a non-repeated $\sigma(t)$) comes from $A\boldsymbol{v} = \sigma\boldsymbol{u}$:

$$U^{\mathrm{T}}AV = \Sigma \qquad \boldsymbol{u}^{\mathrm{T}}(t)\,A(t)\,\boldsymbol{v}(t) = \boldsymbol{u}^{\mathrm{T}}(t)\,\sigma(t)\,\boldsymbol{u}(t) = \sigma(t). \tag{7}$$

The derivative of the left side has three terms from the product rule, as in (5). The first and third terms are zero because $A\boldsymbol{v} = \sigma\boldsymbol{u}$ and $A^{\mathrm{T}}\boldsymbol{u} = \sigma\boldsymbol{v}$ and $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1$. You will see the derivatives of $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}$ and $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$:

$$\frac{d\boldsymbol{u}^{\mathrm{T}}}{dt}\,A(t)\,\boldsymbol{v}(t) = \sigma(t)\,\frac{d\boldsymbol{u}^{\mathrm{T}}}{dt}\,\boldsymbol{u}(t) = \mathbf{0} \quad\text{and}\quad \boldsymbol{u}^{\mathrm{T}}(t)\,A(t)\frac{d\boldsymbol{v}}{dt} = \sigma(t)\,\boldsymbol{v}^{\mathrm{T}}(t)\,\frac{d\boldsymbol{v}}{dt} = \mathbf{0}. \tag{8}$$

The third term from the product rule for $\boldsymbol{u}^{\mathrm{T}}A\boldsymbol{v}$ gives the formula for $d\sigma/dt$:

$$\textbf{Derivative of a Singular Value} \qquad \boxed{\boldsymbol{u}^{\mathrm{T}}(t)\,\frac{dA}{dt}\,\boldsymbol{v}(t) = \frac{d\sigma}{dt}} \tag{9}$$

When $A(t)$ is symmetric positive definite, $\boldsymbol{\sigma}(t) = \boldsymbol{\lambda}(t)$ and $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{y} = \boldsymbol{x}$ in (4) and (9).

Note First derivatives of *eigenvectors* go with second derivatives of eigenvalues—not so easy. The Davis-Kahan bound on the angle θ between unit eigenvectors of S and $S+T$ is $\sin\theta \le ||T||/d$ (d is the smallest distance from the eigenvalue of $S+T$ to all other eigenvalues of S). Tighter bounds that use the structure of S and T are highly valuable for applications to stochastic gradient descent (see Eldridge, Belkin, and Wang).

C. Davis and W. M. Kahan, *Some new bounds on perturbation of subspaces*, Bull. Amer. Math. Soc. **75** (1969) 863 – 868.

J. Eldridge, M. Belkin, and Y. Wang, *Unperturbed: Spectral analysis beyond Davis-Kahan*, arXiv: 1706.06516v1, 20 Jun 2017.

A Graphical Explanation of Interlacing

This page owes everything to Professor Raj Rao Nadakuditi of the University of Michigan. In his visits to MIT, he explained the theory and its applications to the 18.065 class. His **OptShrink** software to find low rank signals is described in *IEEE Transactions on Information Theory* **60** (May 2014) 3002 – 3018.

What is the change in the $\boldsymbol{\lambda}$'s, when a low rank matrix $\theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ is added to a full rank symmetric matrix S? We are thinking of S as noise and $\theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ as the rank one signal. How are the eigenvalues of S affected by adding that signal?

Let me make clear that *all* the eigenvalues of S can be changed by adding $\theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$, not just one or two. But we will see that only one or two have changes of order θ. This makes them easy to find. If our vectors represent videos, and $\theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ represents a light turned on or off during filming (a rank-one signal), we will see the effect on the λ's.

Start with an eigenvalue z and its eigenvector $\boldsymbol{v}$ of the new matrix $S + \theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$:

$$(S + \theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})\boldsymbol{v} = z\boldsymbol{v}. \tag{10}$$

Rewrite that equation as

$$(zI - S)\boldsymbol{v} = \theta\boldsymbol{u}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}) \quad\text{or}\quad \boldsymbol{v} = (zI - S)^{-1}\theta\boldsymbol{u}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}). \tag{11}$$

Multiply by $\boldsymbol{u}^{\mathrm{T}}$ and cancel the common factor $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}$. This removes $\boldsymbol{v}$. Then divide by θ. That connects the new eigenvalue z to the change $\theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ in the symmetric matrix S.

$$\boxed{\frac{1}{\theta} = \boldsymbol{u}^{\mathrm{T}}(zI - S)^{-1}\boldsymbol{u}.} \tag{12}$$

To understand this equation, use the eigenvalues and eigenvectors of S. If $S\boldsymbol{q}_k = \lambda_k\boldsymbol{q}_k$ then $(zI - S)\boldsymbol{q}_k = (z - \lambda_k)\boldsymbol{q}_k$ and $(zI - S)^{-1}\boldsymbol{q}_k = \boldsymbol{q}_k/(z - \lambda_k)$:

$$\boldsymbol{u} = \sum c_k\boldsymbol{q}_k \quad \text{leads to} \quad (zI - S)^{-1}\boldsymbol{u} = \sum c_k(zI - S)^{-1}\boldsymbol{q}_k = \sum \frac{c_k\boldsymbol{q}_k}{z - \lambda_k}. \tag{13}$$

Finally equation (12) multiplies $(zI - S)^{-1}\boldsymbol{u}$ by $\boldsymbol{u}^{\mathrm{T}} = \sum c_k\boldsymbol{q}_k^{\mathrm{T}}$. **The result is $1/\theta$.** Remember that the $\boldsymbol{q}$'s are orthogonal unit vectors:

Secular equation

$$\boxed{\frac{1}{\theta} = \boldsymbol{u}^{\mathrm{T}}(zI - S)^{-1}\boldsymbol{u} = \sum_{k=1}^{n} \frac{c_k^2}{z - \lambda_k}} \tag{14}$$

We can graph the left side and right side. The left side is constant, the right side blows up at each eigenvalue $z = \lambda_k$ of S. The two sides are equal at the n points $z_1, \ldots, z_n$ where the flat $1/\theta$ line meets the steep **curves. Those z's are the n eigenvalues of $S + \theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$. The graph shows that each z_i is above λ_i and below λ_{i-1} : Interlacing.**

The top eigenvalue z_1 is most likely above λ_1. The z's will increase as θ increases and the $1/\theta$ line moves down.

Of course the z's depend on the vector $\boldsymbol{u}$ in the signal (as well as θ). If $\boldsymbol{u}$ happened to be also an eigenvector of S, then its eigenvalue λ_k would increase by exactly θ. All other eigenvalues would stay the same. It is much more likely that each eigenvalue λ_k moves up a little to z_k. **The point of the graph is that z_k doesn't go beyond λ_{k-1}.**

Figure III.1: Eigenvalues z_i of $S + \theta\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ where the $\dfrac{1}{\theta}$ line meets the curves in (14).

The Largest Eigenvalue of $S + T$

The largest eigenvalue of a symmetric matrix S is the maximum value of $\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. This statement applies also to T (still symmetric). Right away we know about the largest eigenvalue of $S+T$.

$$\boxed{\lambda_{\max}(S+T) \le \lambda_{\max}(S) + \lambda_{\max}(T)} \tag{15}$$

The left side is the maximum value of $\dfrac{\boldsymbol{x}^{\mathrm{T}}(S+T)\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$. That maximum is reached at an eigenvector $\boldsymbol{v}$ of $S+T$:

$$\lambda_{S+T} = \frac{\boldsymbol{v}^{\mathrm{T}}(S+T)\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} = \frac{\boldsymbol{v}^{\mathrm{T}}S\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} + \frac{\boldsymbol{v}^{\mathrm{T}}T\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \le \max\frac{\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} + \max\frac{\boldsymbol{x}^{\mathrm{T}}T\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \lambda_S + \lambda_T.$$

The eigenvector $\boldsymbol{v}$ of $S+T$ maximizes that first ratio. But it probably doesn't maximize the last two. Therefore $\lambda_S + \lambda_T$ can only increase beyond λ_{S+T}.

This shows that a maximum principle is convenient. So is a minimum principle for the *smallest eigenvalue*. There we expect $\lambda_{\min}(S+T) \ge \lambda_{\min}(S) + \lambda_{\min}(T)$. The same reasoning will prove it—the separate minimum principles for S and T will bring us lower than $\lambda_{\min}(S+T)$. Or we can apply the maximum principle to $-S$ and $-T$.

The difficulties come for the in-between eigenvalues λ_2 to λ_{n-1}. Their eigenvectors are **saddle points** of the function $R(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. The derivatives of $R(\boldsymbol{x})$ are all zero at the eigenvectors $\boldsymbol{q}_2$ to $\boldsymbol{q}_{n-1}$. But the matrix of second derivatives of R is indefinite (plus and minus eigenvalues) at these saddle points. That makes the eigenvalues hard to estimate and hard to calculate.

We now give attention to the saddle points. One reason is their possible appearance in the algorithms of deep learning. We need some experience with them, in the basic problem of eigenvalues. Saddle points also appear when there are constraints. If possible we want to connect them to *maxima of minima* or to *minima of maxima*.

Those ideas lead to the best possible bound for each eigenvalue of $S+T$:

$$\textbf{Weyl upper bounds} \qquad \lambda_{i+j-1}(S+T) \le \lambda_i(S) + \lambda_j(T). \tag{16}$$

Saddle Points from Lagrange Multipliers

Compared to saddle points, computing a maximum or minimum of $F(\boldsymbol{x})$ is relatively easy. When we have an approximate solution $\widehat{\boldsymbol{x}}$, we know that $F(\widehat{\boldsymbol{x}})$ is not below the minimum of F and not above the maximum: *by definition.* But we don't know whether $F(\widehat{\boldsymbol{x}})$ is *above or below* a saddle point value. Similarly, the matrix $H(\boldsymbol{x})$ of second derivatives of F is positive definite (or semidefinite) at a minimum and negative definite at a maximum.

The second derivative matrix H at a saddle point is symmetric but indefinite.

H has both positive and negative eigenvalues—this makes saddle points more difficult. The conjugate gradient method is not usually available to find saddle points of $\boldsymbol{x}^{\mathrm{T}}H\boldsymbol{x}$.

Lagrange is responsible for a lot of saddle point problems. We start by minimizing a positive definite energy $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}$, but there are m constraints $A\boldsymbol{x} = \boldsymbol{b}$ on the solution. Those constraints are multiplied by new unknowns $\lambda_1, \ldots, \lambda_m$ (the Lagrange multipliers) and they are built into the Lagrangian function:

$$\textbf{Lagrangian} \qquad L(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x} + \boldsymbol{\lambda}^{\mathrm{T}}(A\boldsymbol{x} - \boldsymbol{b}).$$

The $m + n$ equations $\partial L/\partial \boldsymbol{x} = \mathbf{0}$ and $\partial L/\partial \boldsymbol{\lambda} = \mathbf{0}$ produce an indefinite block matrix:

$$\begin{bmatrix} \partial L/\partial \boldsymbol{x} \\ \partial L/\partial \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} S\boldsymbol{x} + A^{\mathrm{T}}\boldsymbol{\lambda} \\ A\boldsymbol{x} - \boldsymbol{b} \end{bmatrix} \quad \text{and} \quad H\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} S & A^{\mathrm{T}} \\ A & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{b} \end{bmatrix} \tag{17}$$

A small example would be $H = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ with negative determinant -1: its eigenvalues have opposite signs. The Problem Set confirms that this **"KKT matrix"** in equation (17) is *indefinite*. The solution $(\boldsymbol{x}, \boldsymbol{\lambda})$ is a saddle point of Lagrange's function L.

Saddle Points from Rayleigh Quotients

The maximum and minimum of the Rayleigh quotient $R(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ are λ_1 and λ_n:

$$\textbf{Maximum} \ \frac{\boldsymbol{q}_1^{\mathrm{T}}S\boldsymbol{q}_1}{\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{q}_1} = \boldsymbol{q}_1^{\mathrm{T}}\lambda_1\boldsymbol{q}_1 = \boldsymbol{\lambda}_1 \qquad \textbf{Minimum} \ \frac{\boldsymbol{q}_n^{\mathrm{T}}S\boldsymbol{q}_n}{\boldsymbol{q}_n^{\mathrm{T}}\boldsymbol{q}_n} = \boldsymbol{q}_n^{\mathrm{T}}\lambda_n\boldsymbol{q}_n = \boldsymbol{\lambda}_n$$

Our question is about the saddle points—the other points where all derivatives of the quotient $R(\boldsymbol{x})$ are zero. We will confirm that **those saddle points occur at the other eigenvectors $\boldsymbol{q}_2$ to $\boldsymbol{q}_{n-1}$ of S. Our goal is to see $\boldsymbol{\lambda}_2$ to $\boldsymbol{\lambda}_{n-1}$ as maxima of minima.** That max-min insight is the key to interlacing.

Notice that the vectors $\boldsymbol{x}$ and $2\boldsymbol{x}$ and $c\boldsymbol{x}$ $(c \neq 0)$ all produce the same quotient R:

$$R(2\boldsymbol{x}) = \frac{(2\boldsymbol{x})^{\mathrm{T}}S(2\boldsymbol{x})}{(2\boldsymbol{x})^{\mathrm{T}}(2\boldsymbol{x})} = \frac{4\,\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}}{4\,\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \frac{\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = R(\boldsymbol{x}).$$

So we only need to consider unit vectors with $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1$. That can become a constraint:

$$\max \frac{\boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} \quad \text{is the same as} \quad \textbf{max}\ \boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x} \ \ \textbf{subject to}\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1. \tag{18}$$

The constraint $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1$ can be handled by one Lagrange multiplier!

$$\textbf{Lagrangian} \qquad L(\boldsymbol{x}, \lambda) = \boldsymbol{x}^{\mathrm{T}}S\boldsymbol{x} - \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - 1). \tag{19}$$

The max-min-saddle points will have $\partial L/\partial \boldsymbol{x} = 0$ and $\partial L/\partial \lambda = 0$ (as in Section I.9):

$$\frac{\partial L}{\partial \boldsymbol{x}} = 2S\boldsymbol{x} - 2\lambda\boldsymbol{x} = \mathbf{0} \quad \text{and} \quad \frac{\partial L}{\partial \lambda} = 1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 0. \tag{20}$$

This says that the unit vector $\boldsymbol{x}$ is an eigenvector with $S\boldsymbol{x} = \lambda\boldsymbol{x}$.

Example **Suppose S is the diagonal matrix with entries 5, 3, 1.** Write x as (u, v, w) :

$$R = \frac{x^{\mathrm{T}} S x}{x^{\mathrm{T}} x} = \frac{5u^2 + 3v^2 + w^2}{u^2 + v^2 + w^2} \text{ has a } \begin{array}{l} \textbf{maximum value 5} \text{ at } x = (1,0,0) \\ \textbf{minimum value 1} \text{ at } x = (0,0,1) \\ \textbf{saddle point value 3} \text{ at } x = (0,1,0) \end{array}$$

By looking at R, you see its maximum of 5 and its minimum of 1. All partial derivatives of $R(u, v, w)$ are zero at those three points $(1, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$. These are eigenvectors of S. $R(x)$ equals the eigenvalues $5, 1, 3$ at those three points.

Maxima and Minima over Subspaces

All the middle eigenvectors $q_2, \ldots, q_{n-1}$ of S are saddle points of the quotient $x^{\mathrm{T}} Sx/x^{\mathrm{T}} x$. The quotient equals $\lambda_2, \ldots, \lambda_{n-1}$ at those eigenvectors. All the middle singular vectors $v_2, \ldots, v_{n-1}$ are saddle points of the growth ratio $||Ax||/||x||$. The ratio equals $\sigma_2, \ldots, \sigma_{n-1}$ at those singular vectors. Those statements are directly connected by the fact that $x^{\mathrm{T}} Sx = x^{\mathrm{T}} A^{\mathrm{T}} Ax = ||Ax||^2$.

But saddle points are more difficult to study than maxima or minima. A function moves both ways, up and down, as you leave a saddle. At a maximum the only movement is down. At a minimum the only movement is up. So the best way to study saddle points is to capture them by a "max-min" or "min-max" principle.

Max-min for λ_2

$$\boxed{\lambda_2 = \max_{\textbf{all 2D spaces } Y} \ \min_{x \textbf{ is in } Y} \frac{x^{\mathrm{T}} Sx}{x^{\mathrm{T}} x}} \tag{21}$$

In the $5, 3, 1$ example, one choice of the 2D subspace Y is all vectors $x = (u, v, 0)$. Those vectors are combinations of q_1 and q_2. Inside this Y, the minimum ratio $x^{\mathrm{T}} Sx/x^{\mathrm{T}} x$ will certainly be $\lambda_2 = 3$. That minimum is at $x = q_2 = (0, 1, 0)$ (we understand minima).

Key point: Every 2D space Y must intersect that 2D space of vectors $x = (0, v, w)$. Those 2D spaces in $\mathbf{R}^3$ will surely meet because $2 + 2 > 3$. For every $x = (0, v, w)$ we definitely know that $x^{\mathrm{T}} Sx/x^{\mathrm{T}} x \le \lambda_2$. So for each Y the minimum in (21) is $\le \lambda_2$.

Conclusion : The maximum possible minimum is λ_2 in (21) and λ_i in (22).

$$\boxed{\lambda_i(S) = \max_{\dim V = i} \ \min_{x \text{ in } V} \frac{x^{\mathrm{T}} Sx}{x^{\mathrm{T}} x} \qquad \sigma_i(A) = \max_{\dim W = i} \ \min_{x \text{ in } W} \frac{||Ax||}{||x||}} \tag{22}$$

For $i = 1$, the spaces V and W are one-dimensional lines. The line V through $x = q_1$ (first eigenvector) makes $x^{\mathrm{T}} Sx/x^{\mathrm{T}} x = \lambda_1$ a maximum. The line W through $x = v_1$ (first singular vector) makes $||Ax||/||x|| = \sigma_1$ a maximum.

For $i = 2$, the spaces V and W are two-dimensional planes. The maximizing V contains the eigenvectors q_1, q_2 and the maximizing W contains the singular vectors v_1, v_2. The minimum over that V is λ_2, the minimum over that W is σ_2. This pattern continues for every i. It produces the Courant-Fischer *max-min principles* in equation (22).

Interlacing and the Weyl Inequalities

For any symmetric matrices S and T, Weyl found bounds on the eigenvalues of $S+T$.

Weyl inequalities $\quad \lambda_{i+j-1}(S+T) \leq \lambda_i(S) + \lambda_j(T) \quad (23)$

$$\lambda_k(S) + \lambda_n(T) \leq \lambda_k(S+T) \leq \lambda_k(S) + \lambda_1(T) \quad (24)$$

The interlacing of the z's that we saw in Figure III.1 is also proved by equation (23). The rank one matrix T is $\theta \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ and its largest eigenvalue is $\lambda_1(T) = \theta$. All of the other eigenvalues $\lambda_j(T)$ are zero. Then for every $j = 2, 3, \ldots$ Weyl's inequality gives $\boldsymbol{\lambda_{i+1}(S+T) \leq \lambda_i(S)}$. Each eigenvalue z_{i+1} of $S+T$ cannot go past the next eigenvalue λ_i of S. And for $j=1$ we have $\boldsymbol{\lambda_1(S+T) \leq \lambda_1(S) + \theta}$: an upper bound on the largest eigenvalue of signal plus noise.

Here is a beautiful interlacing theorem for eigenvalues, when the last row and column of a symmetric matrix S are removed. That leaves a matrix S_{n-1} of size $n-1$.

The $n-1$ eigenvalues α_i of the matrix S_{n-1} interlace the n eigenvalues of S.

The idea of the proof is that removing the last row and column is the same as forcing all vectors to be orthogonal to $(0, \ldots, 0, 1)$. Then the minimum in (22) could move below λ_i. But α_i won't move below λ_{i+1}, because λ_{i+1} allows a free choice with $\dim V = i+1$.

Example
$\lambda_i \geq \alpha_i \geq \lambda_{i+1}$

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \qquad \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \qquad \begin{bmatrix} 2 \end{bmatrix}$$

$$\boldsymbol{\lambda = 3,\ 3,\ 0} \qquad \boldsymbol{\alpha = 3,\ 1} \qquad \boldsymbol{3 > 2 > 1}$$

Interlacing of Singular Values

Suppose A is not square and symmetric—so its singular values are involved. Each column of A represents one frame in a video. We want to identify a rank one signal $\beta \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ hidden in those columns. That signal is obscured by random noise. If a light was turned on or off during the video, the goal is to see when that happened.

This leads us to ask: **How much do the singular values change from A to $A+B$?** Changes in eigenvalues of symmetric matrices are now understood. So we can study $A^{\mathrm{T}}A$ or AA^{T} or this symmetric matrix of size $m+n$ with eigenvalues σ_i and $-\sigma_i$:

$$\begin{bmatrix} 0 & A \\ A^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} = \sigma_i \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} \text{ and } \begin{bmatrix} 0 & A \\ A^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} -\boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} = -\sigma_i \begin{bmatrix} -\boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} \quad (25)$$

Instead we recommend the amazing notes by Terry Tao: **https://terrytao.wordpress.com/ 2010/01/12/254a-notes-3a-eigenvalues-and-sums-of- hermitian-matrices/**

Weyl inequalities $\quad \sigma_{i+j-1}(A+B) \leq \sigma_i(A) + \sigma_j(B) \quad (26)$

$i \leq m \leq n \quad |\sigma_i(A+B) - \sigma_i(A)| \leq ||B|| \quad (27)$

Problem Set III.2

1 A unit vector $\boldsymbol{u}(t)$ describes a point moving around on the unit sphere $||\boldsymbol{u}(t)|| = 1$. Show that the velocity vector $d\boldsymbol{u}/dt$ is orthogonal to the position vector $\boldsymbol{u}(t)$.

2 Suppose you add a positive semidefinite **rank two** matrix to S. What interlacing inequalities will correct the eigenvalues λ of S and α of $S + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$?

3 (a) Find the eigenvalues $\lambda_1(t)$ and $\lambda_2(t)$ of $A = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} + t \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

(b) At $t = 0$, find the eigenvectors of $A(0)$ and verify $\dfrac{d\lambda}{dt} = \boldsymbol{y}^{\mathrm{T}} \dfrac{dA}{dt}\, \boldsymbol{x}$.

(c) Check that the change $A(t) - A(0)$ is positive semidefinite for $t > 0$. Then verify the interlacing law $\lambda_1(t) \geq \lambda_1(0) \geq \lambda_2(t) \geq \lambda_2(0)$.

4 S is a symmetric matrix with eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ and eigenvectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$. Which i of those eigenvectors are a basis for an i-dimensional subspace Y with this property: The minimum of $\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ for $\boldsymbol{x}$ in Y is λ_i.

5 Find the eigenvalues λ of A_3, z of A_2, α of A_1. Show that they are interlacing:

$$A_3 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad A_3 = \begin{bmatrix} 1 \end{bmatrix}$$

6 Suppose D is the diagonal matrix diag $(1, 2, \ldots, n)$ and S is positive definite.

1) Find the derivatives at $t = 0$ of the eigenvalues $\lambda(t)$ of $D + tS$.

2) For a small $t > 0$ show that the λ's interlace the numbers $1, 2, \ldots, n$.

3) For any $t > 0$, find bounds on $\lambda_{\text{min}}\,(D + tS)$ and $\lambda_{\text{max}}\,(D + tS)$.

7 Suppose D is again diag $(1, 2, \ldots, n)$ and A is any n by n matrix.

1) Find the derivatives at $t = 0$ of the singular values $\sigma(t)$ of $D + tA$.

2) What do Weyl's inequalities say about $\sigma_{\text{max}}\,(D + tA)$ and $\sigma_{\text{min}}\,(D + tA)$.

8 (a) Show that every i-dimensional subspace $\boldsymbol{V}$ contains a nonzero vector $\boldsymbol{z}$ that is a combination of $\boldsymbol{q}_i, \boldsymbol{q}_{i+1}, \ldots, \boldsymbol{q}_n$. (Those $\boldsymbol{q}$'s span a space $\boldsymbol{Z}$ of dimension $n - i + 1$. Why does Z intersect $\boldsymbol{V}$?)

(b) Why does that vector $\boldsymbol{z}$ have $\boldsymbol{z}^{\mathrm{T}} S \boldsymbol{z}/\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} \leq \boldsymbol{\lambda}_i$? Then explain Courant-Fischer:

$$\boxed{\lambda_i = \max_{\dim \boldsymbol{V} = i} \quad \min_{\boldsymbol{z} \text{ in } \boldsymbol{V}} \quad \frac{\boldsymbol{z}^{\mathrm{T}} S \boldsymbol{z}}{\boldsymbol{z}^{\mathrm{T}} \boldsymbol{z}}}$$

Problem Set VI.1

1 When is the union of two circular discs a convex set?

2 The text proposes only two ways for the union of two triangles in $\mathbf{R}^2$ to be convex. Is this test correct? What if the triangles are in $\mathbf{R}^3$?

3 The **"convex hull"** of any set S in $\mathbf{R}^n$ is the smallest convex set K that contains S. From the set S, how could you construct its convex hull K?

4 (a) Explain why the intersection $K_1 \cap K_2$ of two convex sets is a convex set.

(b) How does this prove that the maximum F_3 of two convex functions F_1 and F_2 is a convex function? Use the text definition : F is convex when the set of points on and above its graph is convex. What set is above the graph of F_3?

5 Suppose K is convex and $F(x) = 1$ for x in K and $F(x) = 0$ for x not in K. Is F a convex function? What if the 0 and 1 are reversed?

6 From their second derivatives, show that these functions are convex :

(a) Negative entropy $-x \log x$

(b) $\log(e^x + e^y)$

(c) ℓ^p norm $||\boldsymbol{x}||_p = (|x_1|^p + |x_2|^p)^{1/p}$, $p \geq 1$

(d) $\lambda\text{max}\,(S)$

(e) Rayleigh quotient $R(\boldsymbol{x}) = \dfrac{\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}}$ for positive definite S

7 This chapter includes statements of the form $\min\limits_{x}\, \max\limits_{y}\, K(x,y) = \max\limits_{y}\, \min\limits_{x}\, K(x,y)$.

But that is not generally true! Explain this example :

$$\min_{x}\ \max_{y}\ (x+y) \ \text{ and } \ \max_{y}\ \min_{x}\ (x+y) \ \text{ are } +\infty \text{ and } -\infty.$$