MAT-INF4310: Mandatory assignment #4, autumn 2017

To be handed in by November 9., 14:30

You must hand in one (preferably .pdf) file containing your answers as well as commented scripts which actually compile and work.

You must also use "Devilry".

Exercise 1. Let $\mathcal{S} \subset \mathbb{R}^n$ be the simplex

$$\mathcal{S} = \left\{ \mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0 \text{ for } i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} x_i = 1 \right\}.$$

Let also $\mathbf{A}$ be a real $n \times n$ matrix with entries $a_{ij} \geq 0$ which is such that the columns sum to 1, i.e.,

$$\sum_{i=1}^{n} a_{ij} = 1 \text{ for } j = 1, \ldots, n.$$

Such a matrix is sometimes called a *stochastic matrix*.

a) Show that $\mathbf{A}(\mathcal{S}) \subseteq \mathcal{S}$.

b) Show that $\mathbf{A} : \mathcal{S} \to \mathcal{S}$ is continuous in the $\|\cdot\|_1$ norm.

c) Show that if λ is an eigenvalue of $\mathbf{A}$, then $|\lambda| \leq 1$, and that $\mathbf{A}$ has a right eigenpair $(1, \mathbf{w})$ such that $\mathbf{w} \in \mathcal{S}$. Hint: You may use Brouwer's fixed point theorem:

Theorem. *Every continuous function from a convex compact subset K of a Euclidean space to K itself has a fixed point.*

Now we additionally assume that $a_{ij} > 0$ for all i and j.

d) Show that $\mathbf{A}$ maps $\mathcal{S}$ into the interior of $\mathcal{S}$, i.e., the set $\{\mathbf{x} \in \mathcal{S} \mid x_i > 0,\ i = 1, \ldots, n\}$. Conclude that $w_i > 0$ for all i.

e) Show that $A : \mathcal{S} \to \mathcal{S}$ is a contraction in the $\|\cdot\|_1$ norm, i.e., show that for $\mathbf{x}$, $\mathbf{y}$ in $\mathcal{S}$ $\mathbf{x} \neq \mathbf{y}$,

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_1 < \|\mathbf{x} - \mathbf{y}\|_1 .$$

Conclude that the geometric multiplicity of the eigenvalue 1 is one.

The "Page-Rank" algorithm. This algorithm was the initial reason for Google's success. It is a method to determine the "popularity" of each web page, so that a search will list the most popular pages first.

Assume we have a web with n web pages numbered from 1 to n. Each page may contain links to any of the other pages (links to itself do not count). Let ℓ_j be the total number of outgoing links from web page j. According to the founders of Google, Sergey Brin and Larry Page (sic!), the popularity of web page i; p_i, satisfies the equation

$$p_i = \sum_{j \in B_i} \frac{1}{\ell_j} p_j,$$

where

$$B_i = \{\, j \mid \text{there is a link from page } j \text{ to page } i \}\,.$$

The reasoning behind this is that each page gets its popularity by having other pages link to it. Each of these pages then shares its own popularity equally among its outgoing links. Thus incoming links from popular pages with few outgoing links contribute more to your popularity than incoming links from pages with many outgoing links and low popularity.

Let $\mathbf{B}$ be the matrix with entries b_{ij} such that

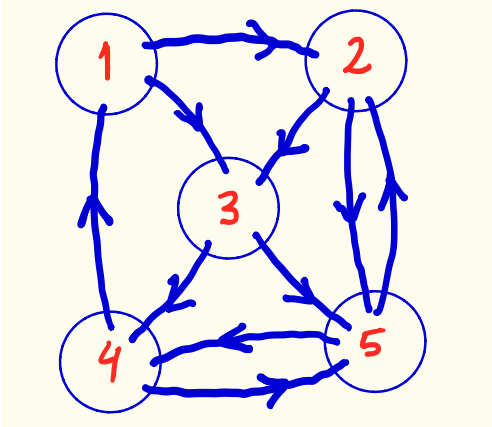
$$b_{ij} = \begin{cases} \frac{1}{\ell_j} & \text{if there is a link from page } j \text{ to page } i, \\ 0 & \text{otherwise,} \end{cases}$$

and let the popularity vector $\mathbf{p} = (p_1, \ldots, p_n)^T$. Since we aim to determine the relative popularity of pages, we normalize $\mathbf{p}$ so that $\sum_i p_i = 1$. Hence the popularity vector satisfies

$$\mathbf{Bp} = \mathbf{p}.$$

Exercise 2.

a) Explain why there always is a popularity vector if all pages have at least one outgoing link, and find the popularity vector of the system of web pages in the figure below. Arrows indicate links.



For the real world wide web, there are roughly 5 billion web pages, so finding an eigenvector with eigenvalue 1 can be laborious. The matrix $\mathbf{B}$ will be sparse, as there are only about 10 links on the average web page, thus the average column will have about 10 entries that are not zero. This is good news when we want to multiply $\mathbf{B}$ with a vector. A fast way of finding an eigenvector with eigenvalue 1 is to start with a $\mathbf{p}^0 \in \mathcal{S}$ and define $\mathbf{p}^{k+1} = \mathbf{B}\mathbf{p}^k$ for $k = 0, 1, 2, \ldots$. However, since $b_{ij} = 0$ for many is and js, we do not know if this converges. Also, if there are pages with incoming links, but no outgoing links, an entire column of $\mathbf{B}$ will be zero, and $\mathbf{B}$ is not stochastic. We first modify $\mathbf{B}$ so that it is stochastic, by "creating weak outgoing links" from pages that have none. If $b_{ij} = 0$ for $i = 1, \ldots, n$, set $x_{ij} = 1/n$ for all i. We write this as

$$\mathbf{X} = \mathbf{B} + \frac{1}{n}\mathbf{C}, \text{ where } c_{ij} = \begin{cases} 1 & \text{if } \mathbf{B}_{\cdot,j} = \mathbf{0}, \\ 0 & \text{otherwise.} \end{cases}$$

In order to create a stochastic matrix for which the iteration is guaranteed to converge, we must modify $\mathbf{X}$ further so that the matrix we use in the iteration has strictly positive entries. To this end, let $\mathbf{1}$ be the $n \times n$ matrix with a all entries equal to 1, and let α be a real number in $(0, 1)$. Set $\mathbf{A} = \alpha\mathbf{X} + (1 - \alpha)\frac{1}{n}\mathbf{1}$. Then $\mathbf{A}$ is stochastic with $a_{ij} > 0$, and the sequence $\left\{\mathbf{p}^k\right\}_{k=1}^{\infty}$ defined by

$$\mathbf{p}^0 \in \mathcal{S}, \ \mathbf{p}^{k+1} = \mathbf{A}\mathbf{p}^k, \ k \geq 0,$$

will converge to a unique $\mathbf{p} \in \mathcal{S}$ which solves $\mathbf{A}\mathbf{p} = \mathbf{p}$. Brin & Page claimed that $\alpha = 0.85$ gives satisfactory results, but it would seem that setting α very close to 1 would produce results that correspond more closely to the true popularity.

b) Knowing that the second largest absolute value of the eigenvalues of $\mathbf{A}$ is α, explain why the founding fathers of Google chose such a "small" α.

The above iteration reads

$$\mathbf{p}^{k+1} = \alpha\mathbf{B}\mathbf{p}^k + \frac{\alpha}{n}\mathbf{C}\mathbf{p}^k + \frac{1 - \alpha}{n}\mathbf{1}\mathbf{p}^k.$$

c) Since n is about 5 billion, one needs efficient matrix vector multiplications. Formulate efficient methods for the products $\mathbf{C}\mathbf{p}$ and $\mathbf{1}\mathbf{p}$.

d) Implement the page-rank algorithm (try to utilize the sparcity of $\mathbf{B}$), and test it on the web stored here. Each line in the file contains a pair of integers i and j, indicating a link from page i to page j. The web has 10 000 pages, so i and j are between 1 and 10 000. List the 10 most popular pages and their popularity.