MATRIX COMPUTATIONS: HOMEWORK 4 (November 21, 2024)

PageRank algorithm

Google Problem describes how Web pages can be meaningfully ranked by a search engine. There are n pages (nodes of a digraph) connected with outgoing and incoming links. Let n_j be the number of outgoing links of page j. The probability of the transition from the page j to the page i is

$$a_{ij} = \begin{cases} \frac{1}{n_j}, & \text{if link } j \to i \text{ exists,} \\ 0, & \text{otherwise.} \end{cases}$$

Then, x_i - score of i-th page of the Web, $i = 1, \ldots, n$ can be defined as follows:

$$x_i = \sum_{j=1}^{n} a_{ij} x_j.$$

The latter says that the rating of a page is the aggregated rating of all its incoming pages multiplied by corresponding transition probabilities. Since the relative ranking matters, we arrive at a Google problem of finding a vector $x^* \in \mathbb{R}^n$ such that

$$Ax^* = x^*, \quad e^\top x^* = 1, \quad x^* \geq 0, \tag{1}$$

where $A = (a_{ij})$, $e = (1, \ldots, 1)^\top$. Note that A is column stochastic, i.e. $a_{ij} \geq 0$ and $e^\top A = e^\top$. Standard results about stochastic matrices show that Google Problem (1) admits a solution (or we can use the duality theorem of linear programming to show this).

Exercise A: Power Method

In order to solve the Google Problem, we apply the iterative Power Method:

$$x^{k+1} := Ax^k, \quad x^0 \in \Delta = \left\{ x \in \mathbb{R}^n \mid e^\top x = 1, \quad x \geq 0 \right\}.$$

(A1) Show that $x_k \in \Delta$ for all $k \geq 0$.

For the following we make the following assumption.

Assumption 1. The matrix A has at least one positive row, i.e., a row whose elements are strictly positive.

We intend to show that the Power Method converges to the (unique) solution of the Google Problem under this assumption.

Let us introduce the following notations. Set the minimal entry of the i-th row of A as $r_i = \min\limits_{1 \leq j \leq n} a_{ij}$. Denote $r = (r_1, \ldots, r_n)$, $\alpha = e^\top r$, $x^0 = \frac{1}{\alpha} r$.

(A2) Show that $r \neq 0$, $x^0 \in \Delta$, and $\alpha \in (0, 1]$.

(A3) Show that the following representation of A is valid:

$$A = (1 - \alpha)\bar{A} + \alpha x^0 e^\top,$$

where $\bar{A}$ is a column stochastic matrix, i.e. $\bar{a}_{ij} \geq 0$ and $e^\top \bar{A} = e^\top$.

(A4) Using A3, show that for all $h \in \mathbb{R}^n$ with $e^\top h = 0$ it holds:

$$\|Ah\|_1 \leq (1-\alpha)\|h\|_1,$$

where $\|y\|_1 = \sum_{i=1}^n |y_i|$ is the 1-norm.

(A5) Show that the solution of Google Problem (1) is unique.

(A6) Using A4, show that the iterates of the Power Method satisfy the estimate

$$\|x^k - x^*\|_1 \leq (1-\alpha)^k \|x^0 - x^*\|_1,$$

where x^* is a solution of the Google problem. Conclude that the Power Method converges to the unique solution of the Google Problem with the rate $(1-\alpha) \in [0,1)$.

Exercise B: PageRank

Let us describe how the Google Problem is solved by the PageRank originally proposed in S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst., 1998. Since the matrix A may not satisfy Assumption 1—specifically, because A may not have positive rows—the Power Method is applied to its perturbed version:

$$A_\beta = (1-\beta)A + \beta M,$$

where $M = (m_{ij})$ with $m_{ij} = \frac{1}{n}$, $i,j = 1,\ldots,n$, and $\beta \in (0,1)$ is a chosen parameter. As a justification for the replacement of A with A_β in the Google Problem, the model of surfer in the Web is proposed. Namely, a surfer follows links with probability $1-\beta$ and jumps to an arbitrary page with probability β.

(B1) Show that A_β has positive rows.

(B2) Applying **Exercice A**, prove that the Power Method converges to the unique solution of the Google Problem for A_β with the rate $1-\beta$.

Exercise C: Collaboration network

We would not ask you to rank websites on the internet. It became way too large for laptops since 1998. Instead, we will ask you to apply PageRank algorithm to rank researchers that study complex networks. You are given a matrix A in a file "collab.xlsx" that represent collaborations in this community. An entry a_{ij} represent the probability that a random surfer starting from the researcher j will jump to the researcher i according to the following procedure: first he chooses a random paper of j with co-authors and then he chooses a co-author i of that paper at random.

The matrix A correspond to the matrix A of **Exercice B**. Please, construct a matrix A_β for $\beta = 0.25$ and apply Power Method to compute the rank of each researcher. As an answer, list Top5 researchers with their scores.

Practical information

The homework solution should be written in English.
Please submit your solution in a `pdf` file named `Group_XX.pdf`.[1]
As you are in master, we strongly recommend you to write your report in latex.
Deadline for turning in the homework: Wednesday December 11, 2024 (11:59 pm).
It is expected that each group makes the homework individually.
If your group has problems or questions, you are welcome to contact the teaching assistants: julien.calbert@uclouvain.be, guillaume.berger@uclouvain.be.

[1] E.g., `Group_01.pdf`, `Group_02.pdf`, `Group_12.pdf`. Failure to comply with these requirements may result in point penalties.