A Mathematical Introduction to Data Science **Feb 26, 2025**

Homework 4. Random Projections

Instructor: Yuan Yao *Due: in 1 week*

The problem below marked by * is optional with bonus credits. For the experimental problem, include the source codes which are runnable under standard settings. Since there is NO grader assigned for this class, homework will not be graded. But if you would like to submit your exercise, please send your homework to the address (datascience.hw@gmail.com) with a title "CSIC5011: Homework #". I'll read them and give you bonus credits.

1. *SNPs of World-wide Populations:* This dataset contains a data matrix $X \in \mathbb{R}^{n\times p}$ of about $p = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $n = 1064$ rows of peoples around the world (but there are 21 rows mostly with missing values). Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

   `https://drive.google.com/file/d/1KMLPEG91mnzdK2pUlq2BkjOx2BsaZy9s/view?usp=sharing`

   which is big (151MB in zip and 2GB original txt). Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and`ind2` such that $X(\texttt{ind1}, \texttt{ind2})$ removes all missing values.

   `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/HGDP_region.mat`

   Another cleaned dataset is due to Quanhua MU and Yoonhee Nam:

   - Genotyped data of the 1043 (n) subjects. 0(AA), 1(AC), 2(CC). Missing values are removed, only autosomal SNPs were selected ($p \approx 400K$). Google drive link: `https://drive.google.com/file/d/1a9I8_akfCMHBRrPMdnWkjyL9fKcQbJJq/view?usp=sharing`
   - Sample Information of 1043 subjects. Google drive link: `https://drive.google.com/file/d/11Q-8B57WDQnrIV92b-h_WLqDGviiYsm2/view?usp=sharing`

   A good reference for this data can be the following paper in Science,

   `http://www.sciencemag.org/content/319/5866/1100.abstract`

   Explore the genetic variation of those persons with their geographic variations, by MDS/PCA. Since p is big, explore random projections for dimensionality reduction.

2. *Phase Transition in Compressed Sensing:* Let $A \in \mathbb{R}^{n\times d}$ be a Gaussian random matrix, *i.e.* $A_{ij} \sim \mathcal{N}(0, 1)$. In the following experiments, fix $d = 20$. For each $n = 1, \ldots, d$, and each $k = 1, \ldots, d$, repeat the following procedure 50 times:

(a) Construct a sparse vector $x_0 \in \mathbb{R}^d$ with k nonzero entries. The locations of the nonzero entries are selected at random and each nonzero equals ± 1 with equal probability;

(b) Draw a standard Gaussian random matrix $A \in \mathbb{R}^{n \times d}$, and set $b = Ax_0$;

(c) Solve the following linear programming problem to obtain an optimal point $\hat{x}$,

$$\begin{aligned} \min_x \quad & \|x\|_1 := \sum |x_i| \\ s.t. \quad & Ax = b, \end{aligned}$$

for example, matlab toolbox `cvx` can be an easy solver;

(d) Declare success if $\|\hat{x} - x_0\| \leq 10^{-3}$;

After repeating 50 times, compute the success probability $p(n, k)$; draw a figure with x-axis for k and y-axis for n, to visualize the success probability. For example, matlab command `imagesc(p)` can be a choice.

Can you try to give an analysis of the phenomenon observed? The following paper by Tropp et al. may give you a good starting point to think.

- Dennis Amelunxen, Martin Lotz, Michael B. McCoy, Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. arXiv:1303.6672. URL: `https://arxiv.org/abs/1303.6672`