School of Computer Science and Technology
University of Science and Technology of China

Marked Exercises for
Algorithms for Big Data
2023 Spring

Due 17 April 2023 at 17:59

Exercise 1 *15 points*
Let $\sum_{i=1}^{r} \sigma_i u_i v_i^T$ be the SVD of A, where $A \in \mathbb{R}^{n \times d}$. Show that $|u_1^T A| = \sigma_1$ and $|u_1^T A| = \max_{\|u\|=1} \|u^T A\|$, where $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$ for a vector $x \in \mathbb{R}^d$.

Exercise 2 *25 points*
Let A be an $n \times d$ matrix with SVD such that $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$. Let $x \in \mathbb{R}^d$ be a vector such that $\|x\|_2 = 1$ and $|x^\top v_1| \geq \delta$ for some $\delta > 0$. Suppose that $\sigma_2 < \frac{1}{2}\sigma_1$. Let w be the vector after $k = \log(1/\varepsilon\delta)$ iterations of the power method, namely,

$$w = \frac{(A^\top A)^k x}{\|(A^\top A)^k x\|_2}.$$

Prove that the length of the projection of w onto the line defined by the first singular vector v_1 is at least $1 - \varepsilon$, i.e., $|w \top v_1| \geq 1 - \varepsilon$.

Exercise 3 *20 points*
Let $k < d$. Let $U \in \mathbb{R}^{d \times k}$ be a random matrix such that its (i, j)-th entry is denoted as u_{ij}, where $\{u_{ij}\}$ are independent random variables such that

$$u_{ij} = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

Now we use matrix B as a random projection matrix. That is, for a (row) vector $a \in \mathbb{R}^d$, we map it to

$$f(a) = \frac{1}{\sqrt{k}} aU$$

For each j such that $1 \leq j \leq k$, define $b_j = [f(a)]_j$, i.e., b_j is the j-th entry of $f(a)$.

- What is the expectation $\mathrm{E}[b_j]$?
- What is $\mathrm{E}[b_j^2]$?
- What is $\mathrm{E}[\|f(a)\|^2]$?

Exercise 4 *20 points*
In the class, we have seen an algorithm, denoted by $\mathcal{A}$, for the (c, r)-ANN problem with success probability at least 0.6. That is, upon a queried vertex x such that there exists a point a^* in the set $\mathcal{P}$ with $d(x, a^*) \leq r$, the algorithm $\mathcal{A}$ outputs some $a \in \mathcal{P}$ with $d(x, a) \leq c \cdot r$ with probability at least 0.6.
Let $\delta \in (0, 1)$. Using the above $\mathcal{A}$ as a subroutine, give a new algorithm $\mathcal{B}$ with success probability at least $1 - \delta$. That is, for the above query vertex x, the algorithm $\mathcal{B}$ outputs some $a \in \mathcal{P}$ with $d(x, a) \leq c \cdot r$ with probability at least $1 - \delta$. Your algorithm should use as little query time as possible. Explain the correctness of your algorithm and state its query time, assuming the query time of $\mathcal{A}$ is $T_{\mathcal{A}}$.

Exercise 5 *20 points*
Let $\alpha \in (0, 1]$. Suppose we change the (basic) Morris algorithm to the following:

(a) Initialize $X \leftarrow 0$

(b) For each update, increment X by 1 with probability $\frac{1}{(1+\alpha)^X}$

(c) For a query, output $\tilde{n} = \frac{(1+\alpha)^X - 1}{\alpha}$.

Let X_n denote X in the above algorithm after n updates.

- Calculate $\mathrm{E}[\tilde{n}]$ and upper bound $\mathrm{Var}[\tilde{n}]$.
- Let $\epsilon, \delta \in (0,1)$. Based upon the above algorithm, give a new algorithm such that with probability at least $1-\delta$, it outputs an estimator $\tilde{n}$ such that $|\tilde{n} - n| \leq \epsilon n$. Explain the correctness and the space complexity (i.e., the number of bits) of your algorithm. It suffices to give an algorithm with space complexity that is a polynomial function of $1/\delta$.

Exercise 6 *Bonus 10 points*
Recall that in the class (see Lecture note 7), we have seen one algorithm based on dimension reduction for solving (c, r)-ANN problem.
Let $0 < p \leq \frac{1}{2}$. Prove that for any $x, y \in \{0,1\}^d$, it holds that

$$\Pr[(Ux)_i \neq (Uy)_i] = \frac{1}{2}\left(1 - (1-2p)^{\mathrm{Ham}(x,y)}\right),$$

where U is a $k \times d$ random matrix such that the entries are independently and identically distributed (i.i.d.) as follows:

$$u_{ij} = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p, \end{cases}$$

and all the calculations are in the finite field $GF(2)$ (i.e., addition and multiplication are always modulo 2).
Hint: You may consider to use the following fact: Let $w \in \{0,1\}^d$ be a random vector such that all entries w_i's are i.i.d. and $\Pr[w_i = 1] = \Pr[w_i = 0] = \frac{1}{2}$ for each $i \leq d$. Then $\Pr[w^\top x \neq w^\top y] = \frac{1}{2}$.
