Big Data Algorithms — List of assignments

Big Data Analytics; winter semester, 2020/2021

January 29, 2021

1 Probability Theory

Exercise 1. *Plot cummulative distribution functions and probability mass funtions (or probability density functions) of the following distributions:*

1. $\mathrm{Uni}(a, b)$ — *Uniform distribution on the interval* $[a, b]$.
2. $\mathrm{Uni}(n)$ — *Uniform distribution on the set* $\{1, 2, \ldots, n\}$.
3. $\mathrm{Ber}(p)$ — *Bernoulli distribution, with probability parameter* p.
4. $\mathrm{Bin}(n, p)$ — *Binomial distribution, with probability parameter* p *and* n *trials.*
5. $\mathrm{Geo}(p)$ — *Geometric distribution, with probability parameter* p.
6. $\mathrm{Exp}(\lambda)$ — *Exponential distribution, with intensity parameter* λ.

Exercise 2. *DONE Show that the variance satisfies the formula:*

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \ .$$

Find the analogous formula for unnormalized Skewness (or third central moment $\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]$*).*

Exercise 3. *Calculate variances of the following distributions:*

1. $\mathrm{Uni}(a, b)$ — *Uniform distribution on the interval* $[a, b]$.
2. $\mathrm{Uni}(n)$ — *Uniform distribution on the set* $\{1, 2, \ldots, n\}$.
3. $\mathrm{Ber}(p)$ — *Bernoulli distribution, with probability parameter* p.
4. $\mathrm{Bin}(n, p)$ — *Binomial distribution, with probability parameter* p *and* n *trials.*
5. $\mathrm{Geo}(p)$ — *Geometric distribution, with probability parameter* p.
6. $\mathrm{Exp}(\lambda)$ — *Exponential distribution, with intensity parameter* λ.

Exercise 4. *Imagine that Andrew draws a number in the following way: He tosses a symmetric coin; if the result is a head, then he draws his favourite number* 2*; otherwise he draws a number from* $\mathrm{Exp}\left(\frac{1}{2}\right)$ *distribution. What is the expected value of such the draw?*

Question 1. *What is the Pearson's correlation coefficient?*

2 BDA

Exercise 5. *There is an unsorted list L of distinct numbers. A size of the list is N (some big number). Propose an algorithm, which searches for the third lowest value in L. Calculate a complexity of this algorithm in terms of Landau's $O(.)$ notation.*

Task 1. *DONE Estimate the number of people in Poland that every fixed day they see a black cat in the morning and are later fired from work on that day.*

Question 2. *DONE Why does the concept of universal hash function make no sense?*

Question 3. *DONE What is the meaning of $\Pr_{h \in \mathcal{H}}(\dots)$ in formulas from this lecture?*

Question 4. *DONE Why the family of all functions from Ω to $[n]$ is a bad universal family of hash functions?*

Question 5. *DONE Do you know an algebraic field with 4 elements?*

Question 6. *DONE Show that 2-independence implies universality.*

Question 7. *DONE Show that k+1-independency implies k-independency.*

Question 8. *DONE What is the value $2^{-\ln(2)}$ (show some approximation) and why this number is important for Bloom filters?*

Task 2 (Coupon Collector Problem). *DONE Calculate the expected number of (uniform) draws that is needed to collect all n different items.*

Exercise 6. *DONE What is the expected number of empty urns after throwing n balls into n urns?*

Exercise 7. *DONE What is the expected number of empty urns after throwing $n \ln(n)$ balls into n urns?*

Exercise 8. *We are throwing k balls into n urns with different indices from the set $[n]$. What is the distribution of a maximal nonempty urn index?*

Question 9. *DONE What is a commutative semigroup?*

Question 10. *DONE What are Cartesian products and projection maps? Are they the same as product mappings and algebraic projection mappings?*

Task 3. *Show that if $\otimes$ is a binary operation which is commutative and associative and $\pi \in S_n$ is an arbitrary permutation of size n, then*

$$\bigotimes_{i=1}^{n} x_{\pi(i)} = \bigotimes_{i=1}^{n} x_i \ .$$

Task 4. *Which of the following functions are commutative and associative?*

1. $+ : \mathbb{C}^2 \to \mathbb{C}$ *DONE*
2. $+_2 : \mathbb{Z}_2^2 \to \mathbb{Z}_2$ *(addition modulo 2) DONE*
3. $\cdot : \mathbb{C}^2 \to \mathbb{C}$ *DONE*
4. $\cdot_2 : \mathbb{Z}_2^2 \to \mathbb{Z}_2$ *(multiplication modulo 2) DONE*
5. $\max : \mathbb{R}^2 \to \mathbb{R}$, *where* $\max(a, b)$ *is the not less number amongst a and b. DONE*
6. $\min : \mathbb{R}^2 \to \mathbb{R}$, *where* $\min(a, b)$ *is the not greater number amongst a and b. DONE*
7. $\wedge : \mathbb{N}^2 \to \mathbb{N}$, *where $a \wedge b = a^b$. DONE*
8. $/ : \mathbb{R}_+^2 \to \mathbb{R}_+$ *(division) DONE*
9. $\cup : \mathcal{P}(\Omega)^2 \to \mathcal{P}(\Omega)$ *(union of sets) DONE*

10. $\cap : \mathcal{P}(\Omega)^2 \to \mathcal{P}(\Omega)$ *(intersection of sets)* DONE
11. $\setminus : \mathcal{P}(\Omega)^2 \to \mathcal{P}(\Omega)$ *(difference of sets)* DONE
12. $\Delta : \mathcal{P}(\Omega)^2 \to \mathcal{P}(\Omega)$, *where* $A\Delta B = A\backslash B \cup B\backslash A$ *(symmetric difference).* DONE
13. $\mathrm{fst} : \Omega^2 \to \Omega$, *where* $\mathrm{fst}(a, b) = a$. DONE
14. $\mathrm{Av} : \mathbb{R}^2 \to \mathbb{R}$, *where* $\mathrm{Av}(a, b) = \frac{a+b}{2}$. DONE

Task 5. *Which of the following operations are amenable for MapReduce algorithm which uses the Compose arrangement?*

1. *Count of elements* DONE
2. *Count of odd elements* DONE
3. *Max, Min, Range of elements* DONE
4. *Mode,* DONE *Median of elements*
5. *Sum of elements* DONE
6. *Sum modulo 2 of elements* DONE
7. *Sum even elements* DONE
8. *Arithmetic mean of elements* DONE
9. *Product of elements* DONE
10. *Geometric mean of elements* DONE
11. *Harmonic mean of elements* DONE
12. *Sum of squares of elements* DONE
13. *Sum of cubes of elements* DONE
14. *Sum of square roots of elements* DONE
15. *Empirical second moment of elements* DONE
16. *Empirical variance of elements (second central moment)* DONE
17. *Empirical standard deviation of elements* DONE
18. *Empirical third central moment* DONE
19. *Empirical skewness* DONE
20. *Empirical kurtosis* $\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}}\right)\right]$ DONE
21. *Sum of* $\exp(.)$ *applied to elements (sum of product mappings* $\exp(.)$ *on elements)* DONE
22. *Empirical characteristic function of the distribution of elements* $\varphi_X(t) = \mathbb{E}\left[\exp\{itX\}\right]$ *(Fourier transform)* DONE
23. *Empirical Laplace transform of the distribution of elements* $\mathcal{L}(X)(\lambda) = \mathbb{E}\left[\exp\{\lambda X\}\right]$ DONE
24. *Set of elements (without repetitions)* DONE
25. *Count of distinct elements* DONE

Exercise 9 (Maciej Gębala's exercise). *Consider the hash function is given by the formula* $h(x) = x(mod\, 21)$. *We apply it to the numbers divisible by a certain constant* c. *For which constants* c, h *is the proper hash function, i.e for which constants* c *it can be expected that the distribution of bucket loading* $\{0, ..., 20\}$ *will be uniform?*

Exercise 10 (Macej Gębala's exercise). *Find the formula for the order of element* $k \in \{0, ..., n-1\}$ *in a group* $\mathbb{Z}_n$. *What is the relationship between this problem and the previous one?*

Exercise 11 (Maciej Gębala's exercise). *Design the MapReduce algorithm, which determines joining of two relations defined as* $R(A, B, C)$ *and* $S(X, Y, Z)$ *schemes, according to the* $B = X$ *and* $C = Y$ *connection. In other words, find the following set:*

$$\{(A, Z) : (\exists\, B, C) R(A, B, C) \wedge S(B, C, Z)\}\ .$$

Task 6 (Maciej Gębala's exercise). *Let* $F : (\mathbb{N} \times \mathbb{R})^2 \to (\mathbb{N} \times \mathbb{R})$ *be a function specified by the formula*

$$F([c_1, x_1], [c_2, x_2]) = \left[c_1 + c_2, \frac{c_1 x_1 + c_2 x_2}{c_1 + c_2}\right]\ .$$

1. *Show that* F *is associative and commutative.*
2. *Let us denote* $x \oplus y = F(x, y)$. *Find a compact formula for* $[c_1, x_1] \oplus [c_2, x_2] \oplus \ldots \oplus [c_n, x_n]$.
3. *Use this property of the function* F *to design a Combiner in order to determine the mean and variance.*

Exercise 12 (Maciej Gębala's exercise). *Use the MapReduce method to designate all anagrams that appear in a text file.*

Question 11. DONE *What is a simple graph? What is a directed graph? What is a neighbour of a vertex? What is a degree of a vertex? What is an adjacency matrix?*

Exercise 13. DONE *Use the MapReduce method to designate degrees of all vertices of a graph.*

Exercise 14. DONE *Use the MapReduce method to designate all neighbours of a given vertex in a graph.*

Exercise 15. *Use the MapReduce method to designate all two-steps neighbours of a given vertex* v *in a graph (do not count vertex* v*!).*

Question 12. *What is a metric function and metric space? How to define a metric on a graph?*

Question 13. DONE *How to multiply block matrices? For instance how to calculate*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} ?$$

Question 14. DONE *How to exponentiate matrices?*
For instance how to calculate

$$\begin{bmatrix} 4 & 2 & 3 \\ 1 & -1 & -3 \\ -1 & -2 & 0 \end{bmatrix}^{10} ?$$

Exercise 16. DONE *Imagine that we are using Vitter's R algorithm with reservoirs of size* 1 *independently* 5 *times n the same stream* S *in order to obtain the reservoir sample of size* 5 *with eventual repetitions. What is the probability that the repetition occurs after reading* n *elements of the stream?*

Task 7. DONE *Assume that we have reservoir* R *of a fixed size* r: $[R[1], R[2], \ldots, R[r]]$, *initially containing nulls and we are observing a stream* S:$\{S[1], S[2], \ldots\}$. *We initialize the reservoir by*

```
Initialize(S,R,n,r){\\
    n:=0;\\
        onRead(x,S){ \*read the element x from the stream S*\ \\
            n++;\\
            If (n<=r) {\\
                R[n]:=x;\\
            }
            Else {Return;}
        }
}
```

Consider a sampling algorithm R:

```
Update(S,R,n,r) {\\
    onRead(x,S){\\
        n++;\\
        If(n>r && rand() < p(n)) {\\
            pos:=randInt(r);
            R[pos]:=x;
        }
    }
}
```

1. *What does Initialize do?*
2. *What does Update do?*
3. *What does it mean that a sample of size r is uniformly distributed among n elements?*
4. *What $p(n)$ should be in order to provide a uniform reservoir sample R of a stream S at moment n?*

Exercise 17. DONE *Vitter's R algorithm with sample of size 1, which updated a sample at moment n, will provide the next update after reading exactly L elements, where L is equal with respect to the distribution with $\left\lceil \frac{nu}{1-u} \right\rceil$, where $u \sim \mathrm{Uni}\,(0,1)$.[Lectures]*
Find sensible bounds of expected value $\mathbb{E}(L)$. How to use those facts in order to adjust the algorithm?

Exercise 18. DONE *What is the distribution of L from the previous Exercise?*

Question 15. DONE *What are the first terms of Taylor series of $\ln(1+x)$ at $x=0$?*
What is the possible improvement of the estimator of number of distinct elements from the lecture:

$$\hat{m} = N \ln\left(\frac{N}{u}\right) \approx \frac{\ln\left(\frac{u}{N}\right)}{\ln\left(1-\frac{1}{N}\right)} \ ?$$

(N is a range of hash function, u is a number of 0 slots in $[n]$)

Question 16. DONE *Let (LC_i) be a sequence of independent linear counters (estimators $\hat{m}$) of the same set. Why the results obtain as*

$$\frac{1}{k}\sum_{i=1}^{k} \mathrm{LC}_i$$

is more precise?

Exercise 19. DONE *Let L be a random variable, which is a number of empty slots in $[N]$ after using a linear count on a set of m distinct elements. Then*

$$\mathbb{E}\left[L\right] = \left(1-\frac{1}{N}\right)^m \ .$$

What is a $\mathrm{Var}[L]$?

Question 17. *DONE Recall the Chebyshev (—Bienaymé) inequality.*
Calculate

$$\frac{\mathrm{Var}[L]}{(\mathbb{E}\,[L])^2}$$

for the linear counter and use the result together with Chebyshev (—Bienaymé) inequality in order to obtain some restriction of L.
How it affects the estimator $\hat{m}$?

Question 18. *DONE Draw a transition graph of Markov chain associated with standard Morris counter $(C_n, n \in \mathbb{N}_0)$.*
What is a distribution of C_7?

Exercise 20. *From the lecture, we have already know that $\mathbb{E}\left[2^{C_n}\right] = n+1$ for the standard Morris counter. How does it change, when we substitute the update probability from 2^{-C_n} to a^{-C_n} (where $a > 1$)?*

Exercise 21. *DONE What is an unbiased estimator? Provide the unbiased estimator $\hat{n}$ of n, which uses the standard Morris counter. Calculate $\mathrm{Var}(\hat{n})$.*

Exercise 22. *DONE What is a confidence interval? Attain a neat confidence interval for the estimator $\hat{n}$. Hint: Use Chebyshev—Bienaymé inequality.*

Question 19. *DONE Why Geometric Counter is used to establish the approximate value of distinct elements in a set?*
Why the improvements of GC are relying to "LogLog"?

Question 20. *What is the biggest value of the upgraded MaxGeometric counter, using one hash function. ($L - M + 1$ from the lecture, which is obtain, when all bits of s_2 are zeros.)*

Exercise 23. *Prove Cauchy's inequalities of means:*

$$\frac{\sum_{i=1}^{n} a_i}{n} \geq \sqrt[n]{\prod_{i=1}^{n} a_i} \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}\ .$$

Question 21. *DONE What is the approximate value of*

$$\varphi = \frac{\exp\{\gamma\}}{\sqrt{2}} \cdot \frac{2}{3} \prod_{n=1}^{\infty} \left(\frac{(4n+1)(4n+2)}{4n(4n+3)}\right)^{\epsilon_n}\ , \tag{1}$$

where $\gamma = 0,57721\ldots$ is Euler—Mascheroni constant and ϵ_n is $\{-1,1\}$-Morse—Thue sequence (if $\nu(n)$ is the number of occurances of digit 1 in the binary representation of natural number n, then $\epsilon_n = (-1)^{\nu(n)}$)?
How it relates to HyperLogLog algorithm?

Question 22. *DONE Which version of HyperLogLog is the best?*

Question 23. *DONE What is a metric space? What is a normed space?*

Exercise 24. *DONE Show that the function $d(A,B) = |A \Delta B|$ is a metric over the space of non-empty finite subsets of any fixed set X.*

Exercise 25. *DONE Show that $(\{0,1\}^n, d_H)$, where d_H is the Hamming metric is isomorphic with $(\mathcal{P}(\{1,2,\ldots,n\}), d)$, where $d(A,B) = \|A\Delta B|$. (Indicate an isomorphism and show that it preserves the metric)*

Question 24 (Steinhaus' Theorem)**.** *Let (X,d) be some metric space, $a \in X$ and*

$$\rho(x,y) = \frac{2d(x,y)}{d(a,y) + d(x,y) + d(x,a)}$$

be the Steinhaus distance on X. Show that (X,ρ) is a metric space and calculate $\rho(x,a)$, assuming that $x \neq a$. What is the maximal value of ρ?
What is the correspondence between ρ and the Jaccard distance?

Exercise 26 (Maciej Gębala's exercise). *Let's choose two random m-element subsets A and B from the n-element set X. What is the expected value of Jaccard similarity $J(A, B)$?*

Exercise 27. DONE *Show that algebraic structrure $(S_n, \circ)$ of permutations of size n is a group.*

Exercise 28. DONE *Let $X = \mathbb{R}^n$ be a linear space, $\|.\|$ be a second norm, i.e. $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ and $\circ$ be a standard scalar product in Euclidean space $\mathbb{R}^n$. Show that*

$$(\forall\, x, y \in X)\; x \circ y = \|x\| \cdot \|y\| \cdot \cos(\alpha(x, y)) \;,$$

where $\alpha(x, y)$ denotes the angle between x and y.

Exercise 29. *Let $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$, $\|.\|$ be the second norm and $\circ$ is the standard scalar product. Show that $d(x, y) = 1 - x \circ y$ is not a metric on the hypersphere S^{n-1}.*

Question 25. DONE *During the lecture, it has been proven that if $\alpha(x, y)$ is a smaller of the angles between $x, y \in \mathbb{R}^n$, then $\bar{d}(x, y) = \frac{\alpha(x,y)}{\pi}$ defines a normalized metric on S^{n-1}.*
Let $x, y \in S^{n-1}$. What is the value of $\bar{d}(x, y) + \bar{d}(-y, x - y) + \bar{d}(-x, y - x)$?

Question 26. DONE *In order to provide sketches of similarity of some vectors like documents of words, we would like to generate uniformly some random lines $Ax + By = 0$. How to generate A and B properly?*

Exercise 30. DONE *Rewrite Misra—Gries Algorithm for $L = 1$ in order to obtain Majority Counting algorithm.*

Question 27. DONE *What is the result of Majority Counting algorithm for a stream $s = (\underbrace{a, a, \ldots, a}_{n}, \underbrace{b, b, \ldots, b}_{n}, c)$?*
Consider a probabilistic space of permutations of size $2n + 1$ (i.e. S_{2n+1}). What is the distribution of results of $\pi(s)$, assuming that $\pi \in S_{2n+1}$ is drawn uniformly.

Question 28. DONE *During the lecture, it has been showed that if for $L = 1$, some value occurs more than $\frac{N}{2}$ times (N is a size of a stream), then this key will be obtained by Majority Counting algorithm. Can you provide similar result for other L for Misra—Gries Algorithm?*

Exercise 31. DONE *Prove Markov's inequality: If $X \geqslant 0$ and $\mathbb{E}[X] < \infty$, then for any $a > 0$ the following is satisfied:*

$$\mathbb{P}[X \geqslant a] \leqslant \frac{\mathbb{E}[X]}{a} \;.$$

Task 8. *Consider a Min-Count Sketch algorithm for an element a with ω hash functions for a stream of size N with sketches of size L.*

1. *Prove that*
$$\mathbb{P}\left[\bigwedge_{i=1}^{\omega} fr(a) \leqslant m_i(a) \leqslant fr(a) + \frac{cN}{L}\right] \geqslant 1 - c^{-\omega} \;.$$
2. *Assume that we are searching for accomodation and we browse the offers of types "house", "flat" and "bungalow" on some website with 10000 different offers, 1500 offers of type "house", 3000 offers of type "flat" and 500 offers of type "bungalow". Assume that we use Min-Count Sketch algorithm with $\omega = 10$ and $L = 1000$ in order to find the sum of numbers of offers we are interested in (5000). Find some lower bound for a probability that the sum S provided by the algorithm will have the property $|S - 5000| \leqslant 60$.*
3. *What would happen if the total number of offers were 10^6?*

Question 29. *How to choose uniformly a random point from the ball*

$$B((0, 0), 1) := \{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} \leqslant 1\} \;.?$$

Remark that the the expected value of the length of such the vector should be $\frac{2}{3}$.

Question 30. *Let B be a ball $B(0,1)$, $A \subset B$ and X be unifromly distributed in B. Why and when the beneath property is satisfied:*

$$\mathbb{P}\left[X \in A\right] = \frac{vol(A)}{vol(B)} \ ?$$

Exercise 32. *Define Euler's Gamma function (for $z \in \mathbb{R}_+ \cup \{0\}$):*

$$\Gamma(z) = \int\limits_0^\infty t^{z-1} e^{-t} dt \ .$$

Show that if $n \in \mathbb{N}_0$, then $\Gamma(n+1) = n!$ and $\Gamma(n+\frac{1}{2}) = \sqrt{\pi}\frac{(2n)!}{4^n n!}$.

Question 31. *What is a unitary space? How we can define an orthogonal vector projection onto some linear subspace of $\mathbb{R}^n$?*

Exercise 33. *During the lecture there have been shown that a proportion of the volume of $B_n(1)$ – n-dimensional hyperball of radius 1 – and the volume of $C_n(2)$ — n-dimensional hypercube with the edge of length 2 – tends very fast to 0.*
What is the length of an edge of the cube when $vol(B_n(1)) = vol(C_n(a))$?