Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Class Test 1, Autumn 2023
CS60021: Scalable Data Mining

Full Marks: 40 **Time: 30 mins**

Answer ALL questions. You can use calculators.

1. Write the ADMM updates for the following problem:

$$\min_{x,y} \; x^2 + y^2$$
$$\text{sub. to. } \; x = y$$

You can use λ as the Lagrange multiplier. [10]

2. Write the Nesterov's accelerated stochastic gradient descent updates for the objective function:

$$\min_{x} f(x) = \sum_{i=1}^{m} [b_i(x - a_i)^2]$$

[10]

3. Consider a simple SGD update that has been executed for T iterations:
$w = w - \eta v$ where $E[v] = \nabla f$
f is convex, $||w|| \leq B$ and $|v| \leq \rho$.

Write a bound on the fixed learning rate η (not a function of t) such that the return value $\bar{w} = 1/T \sum w_t$ converges to w^*.

How many iteration would it take achieve an error ($||\bar{w} - w^*||$) of at most 10^{-3}. [10]

■

Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Mid-semester Examination, Autumn 2023
CS60021: Scalable Data Mining

Full Marks: 60 **Time: 2 hrs**

Answer ALL questions. You can use calculators.

1. State whether the following statements are true or false with 1 – 2 sentences of explanations: **[5 x 2 = 10]**

   (a) Join operation in Spark always induces "fat" dependencies.

   (b) Total memory consumed by all mapper output records should fit into the combined main memory of all the computers in the network.

   (c) With an appropriate learning rate, the objective function value for every run of SGD converges.

   (d) Batchnorm layers do not introduce any new parameters to the neural network.

   (e) Fully-distributed ADMM without any central coordinating server can potentially introduce $O(n^2)$ extra parameters, where n is the number of computers.

2. Write a Spark program that takes an RDD `points` as input, where each record is an image (256×256 matrix), and another image `testpoint`. The program outputs list of k-nearest images to `testpoint`. Note that here k is small and hence you can store a list k nearest points as a local variable. You are not allowed to use the Spark built-in sort function, or any other built-in top-k functions. **[10]**

3. State 2 design objectives of HDFS which make it different from other distributed file systems like NFS. Name the components of HDFS. Which component is responsible for initiating self-healing? Explain the self-healing process. **[5]**

4. Which component (class) of Pytorch is responsible for the automatic differentiation property? Consider the following code: **[1+2+7=10]**

```
x = torch.autograd.Variable(torch.tensor([10.0,10.0]),requires_grad=True)
y = torch.autograd.Variable(torch.tensor([5.0,5.0]),requires_grad=True)
z = x+y
u = z.mean()
u.backward()
print(x.grad)
print(y.grad)
```

Write the output of the above program. Draw the forward and backward graphs for the above program with the values of appropriate variables at each node.

5. Derive the ADMM formulation for optimizing from first principles: **[10]**

$$\min_{x_1, x_2} \; (a_1x_1 - b_1)^2 + (a_2x_2 - b_2)^2$$
$$\text{Subject to: } x_1 = x_2$$

such that updates to x_1 and x_2 happen in parallel. Clearly state the dual and consensus variables. Write expressions for primal and dual residuals.

6. Write the updates for the ADAM algorithm for minimizing a loss function of the form $L(w) = \frac{1}{n}\sum_{i=1}^{n} l(w, x_i, y_i)$. Derive the ADAM updates for the least square regression problem for the predictor: $y = w_1x_1 + w_2x_2$. Show 4 updates of the ADAM optimizer when the above model is trained on the following dataset: $(x_1, x_2, y) = \{(1, 1, 5), (-1, 1, 2), (1, -1, 2)\}$ with learning rate of 0.01 starting from $(0, 0)$ and momentum parameter as 0.9. **[10]**

7. What is the purpose of the Stochastic Averaged Gradient (SAG) descent algorithm? Define convergence rates for optimization algorithms and explain the above answer. Write the SAG algorithm. **[2+3=5]**

■

Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
Class Test 2, Autumn 2023
CS60021: Scalable Data Mining

Full Marks: 30 **Time: 45 mins**

Answer ALL questions. You can use calculators.

Question 1

(5 + 5 = 10)

a. Consider the following documents:
D1: ABCDEFGH
D2: ABBCDDEFFGHH
Calculate the Jaccard similarity between these documents using 3-shingles.

b. Write the algorithm for computing the Minhash signatures given a term (shingle) - document matrix M. Clearly describe the input and output.

Question 2

(5 + 5 = 10 marks)

a. Write the algorithm for Count-min sketch.

b. You have designed a Count-Min sketch. However, when returning the estimated frequency for a query, you forgot to return the minimum of the estimates, instead you returned the median (i.e. you return $Median_i A[i, h_i(x)]$). Can you derive a range bound for this modified sketch ?

Question 3

(3 + 7 = 10 marks)

a. What is the problem of set membership in the streaming setting? State a practical application.

b. Describe the algorithm for Cuckoo filter. What is the key difference between the cuckoo filter and th Bloom filter.

Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
End-semester Examination, Autumn 2023
CS60021: Scalable Data Mining

Full Marks: 100 **Time: 3 Hrs**

Answer ALL questions. You can use calculators.

Question 1 State whether following statements are true or false with 1 – 2 sentences of explanations:

(10 x 2 = 20 marks)

a. Bloom filter does not allow deletion to attain zero false negative rate.
b. There is no way to determine whether an element returned by the Majority algorithm is actually a majority element in the stream or not.
c. Both count-sketch and count-min sketch maintain the same sketch. Only the return value changes from one to the other.
d. Output of FM-sketch is independent of the order of items in the stream.
e. The family of hashing functions used for LSH corresponding to the Hamming Distance is the Minhash function.
f. Every member f of a gap-LSH family of hashing functions satisfies the equality $Sim(x,y) = P(f(x) = f(y))$.
g. The expected value of minimum of k real uniform random numbers in the range $[0,1]$ is $\frac{1}{k+1}$.
h. The main disadvantage of the Flajolet Martin sketch over k-minimum value sketch is that it is more memory heavy.
i. Median of means method can be used boost the probability of any approximation scheme for an optimization problem, to provide an $\epsilon - \delta$ approximation.
j. Any set membership data structure, e.g. Bloom filter or Cuckoo Filter can be modified to provide frequency counts for the elements.

Question 2

(7 + 8 + 10 = 25)

a. Given that one is using a b-bit uniformly random hashing function, what is the chance of collision between a pair ? What is the chance of at least one collision as we use b-bit hashing function for a Bloom filter to encode n elements ? Show that for probability of collision to be less than 1/n, it is sufficient to have

$$b \geq 3 \log_2 n$$

b. Show that one needs $O(n \, log \left(\frac{1}{\delta}\right))$ space for storing n elements in a Bloom filter, with false positive rate of at most δ.

c. Consider a stream of users arriving at a website, where each user can be either a male or a female (this information is known). We are interested in maintaining a random sample of k users with the following properties:

i. Each male and each female have equal probability of being selected.

ii. Total number of males and females in the sample are roughly equal.

Design a reservoir sampling scheme for the same. Show that these properties are indeed satisfied.

Question 3

(5 + 5 + 7 + 8 = 25 marks)

a. What are ϵ-heavy hitters in a stream ? Write the Misra-Gries algorithm for finding ϵ-heavy hitters.

b. Consider the stream: 1 1 2 2 3 3 4 4 4. Run the Misra-Gries algorithm with 2 counters, and report the heavy hitters and their counts. Give a permutation of the above stream such that 3 is reported as a heavy hitter.

c. Write the Count-sketch algorithm. Write how count-sketches A_1 and A_2 for the two streams S_1 and S_2 can be combined to get counts for the stream $A_1 \cup A_2$.

d. Show that $O(^1/_{\epsilon^2} \log\left(\frac{1}{\delta}\right) \log(n))$ space is sufficient to provide ϵ-accurate frequencies with probability $1 - \delta$, for n elements.

Question 4

(5 + 10 = 15 marks)

a. Define Submodular Functions, mathematically. What are monotone submodular functions? Consider the function f for a set S, $f(S) = \sum_{x \in S} g(x)$, where $x \in X$ are elements that constitute the sets, and $g: X \rightarrow [0, \infty)$ are positive real functions. Show that f is a monotone submodular function.

b. Write the greedy approximation algorithm for maximization for monotone submodular functions. State and derive the approximation error bound for the above algorithm.

Question 5

(10 + 5 = 15 marks)

a. Consider the following sentences:

S1: *Only he told his wife that he loved her.*
S2: *He told his wife that only he loved her.*
S3: *He told his wife that he loved her only.*

Calculate and show the shingle sets for above sentences, considering words as basic symbols (not characters) and using 2-shingles.

Using the above written shingle sets, calculate the min-hash signatures using hash functions: ***(x + 2) mod 10*** and ***(3x + 2) mod 10***, where x is the index.

Use the minhash signatures to estimate the Jacard similarity between each pair of sentences.

b. What is the minimum number of hash functions n from which if the Jaccard similarity is estimated, you can expect a variance of 0.2 or less ? Explain.