Sublinear Algorihms for Big Data

Exam Problems

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

Problem 1: Modified Chernoff Bound

- Problem: Derive Chernoff Bound #2 from Chernoff Bound #1. Explain your answer in detail.
- (Chernoff Bound #1) Let $\boldsymbol{X}_1 \dots \boldsymbol{X}_{\boldsymbol{t}}$ be independent and identically distributed r.vs with range [0, 1] and expectation μ. Then if $X = \frac{1}{\boldsymbol{t}}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\boldsymbol{t}\mu\delta^2}{3}\right)$$

- (Chernoff Bound #2) Let $\boldsymbol{X}_1 \dots \boldsymbol{X}_{\boldsymbol{t}}$ be independent and identically distributed r.vs with range $[0, \boldsymbol{c}]$ and expectation μ. Then if $X = \frac{1}{\boldsymbol{t}}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\boldsymbol{t}\mu\delta^2}{3\boldsymbol{c}}\right)$$

Problem 2: Modified Chebyshev

- Problem: Derive Chebyshev Inequality #2 from Chebyshev Inequality #1. Explain your answer in detail.
- (Chebyshev Inequality #1) For every $c > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- (Chebyshev Inequality #2) For every $c' > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'\,\mathbb{E}[\boldsymbol{X}]\right] \leq \frac{Var[\boldsymbol{X}]}{(c'\,\mathbb{E}[\boldsymbol{X}])^2}$$

Problem 3: Sparse Recovery Error

- Definition (frequency vector): f_i = frequency of i in the stream = # of occurrences of value i
$$f = \langle f_1, \dots, f_{\boldsymbol{n}} \rangle$$
- Sparse Recovery: Find g such that $||f - g||_1$ is minimized among g's with at most k non-zero entries.
- Definition: $Err^k(f) = \min_{g:||g||_0 \le k} ||f - g||_1$
- Problem: Show that $Err^k(f) = \sum_{i \notin S} |f_i|$ where S are indices of k largest f_i. Explain your answer **formally** and **in detail**.

Problem 4: Approximate Median Value

- Stream: $\boldsymbol{m}$ elements $x_1, \dots x_{\boldsymbol{m}}$ from universe $[\boldsymbol{n}] = \{1, 2, \dots, \boldsymbol{n}\}$.
- $S = \{x_1, \dots, x_m\}$ (all distinct) and let
$$rank(y) = |x \in S : x \leq y|$$
- Median $\boldsymbol{M} = x_i$, where $rank(x_i) = \frac{\boldsymbol{m}}{2} + 1$ ($\boldsymbol{m}$ odd).
- Algorithm: Return $\boldsymbol{y}$ = the median of a sample of size $\boldsymbol{t}$ taken from S (with replacement).
- Problem: Does this algorithm give a 10% approximate value of the median with probability $\geq \frac{2}{3}$, i.e. y such that
$$\boldsymbol{M} - \frac{\boldsymbol{n}}{10} < \boldsymbol{y} < \boldsymbol{M} + \frac{\boldsymbol{n}}{10}$$
if $\boldsymbol{t} = o(\boldsymbol{n})$? Explain your answer.

Problem 5: Lower bound on F_k

- Definition (frequency vector): f_i = frequency of i in the stream = # of occurrences of value i

$$f = \langle f_1, \ldots, f_{\boldsymbol{n}} \rangle$$

- Define (frequency moment): $F_k = \sum_i f_i^k$
- Problem: Show that for all integer $\mathrm{k} \geq 1$ it holds that:

$$F_k \geq n \left(\frac{m}{n}\right)^k$$

- Hint: worst-case when $f_1 = \cdots = f_n = \frac{m}{n}$. Use convexity of $g(x) = x^k$

Problem 6: Approximate MST Weight

- Let G be a weighted graph with weights of all edges being integers between 1 and W.
- Definition: Let n_i be the # of connected components if we remove all edges with weight $> (1+\epsilon)^i$.
- Problem: For some constant $\boldsymbol{C} > 0$ show the following bounds on the weight of the minimum spanning tree $w(MST)$:

$$w(MST) \leq \sum_{i=0}^{\lceil \log_{1+\epsilon} W \rceil} \epsilon\,(1+\epsilon)^i\, n_i$$

$$\leq (1 + \boldsymbol{C}\,\epsilon) w(MST)$$

Evaluation

- Deadline: September 1st , 2014, 23:59 GMT
- Submissions in English: grigory@grigory.us
  - Submission Title (once): Exam + Space + “Your Name”
  - Question Title: Question + Space + “Your Name”
  - Submission format
    - PDF from LaTeX (best)
    - PDF
- You can work in groups (up to 3 people)
  - Each group member lists all others on the submission
- Point system
  - Course Credit = 2 points
  - Problems 1-3 = 0.5 point each
  - Problem 4 = 1 point
  - Problems 5-6 = 2 point each