CSC2420 Fall 2012: Algorithm Design, Analysis and Theory

Allan Borodin

November 23, 2012; Lecture 11

Sublinear space and the streaming model

- Sublinear space has been an important topic in complexity theory since the start of complexity theory. While not as important as the $P = NP$ or $NP = co - NP$ question, there are two fundamental space questions that remain unresolved:
  1. Is $NSPACE(S) = DSPACE(S)$ for $S \geq \log n$?
  2. Is P contained in $DSPACE(\log n)$ or $\cup_k SPACE(\log^k n)$? Equivalently, is P contained in polylogarthmic parallel time.
- Savitch [1969] showed a non deterministic S space bounded TM can be simulated by a deterministic S^2 space bounded machine (for space constructible bounds S).
- Further in what was considered a very surprising result, Immerman [1987] and independently Szelepcsényi [1987] $NSPACE(S) = co - NSPACE(S)$. (Savitch's result was also considered suprising by some researchers when it was announced.)

USTCON vs STCON

We let *USTOC* (resp. *STCON*) denote the problem of deciding if there is a path from some specified source node s to some specified target node t in an unidirected (resp. directed) graph G.

- As previously mentioned the Aleliunas' et al [1979] random walk result showed that *USTCON* is in $RSPACE(\log n)$ and after a sequence of partial results about *USTCON*, Reingold [2008] was eventually able to show that *USTCON* is in $DSPACE(\log n)$
- It remains open if *STCON* is in $RSPACE(\log n)$ (and hence $NSPACE(S) = RSPACE(S)$ or if $RSPACE(S) = DSPACE(S)$.

The streaming model

- In the data stream model, the input is a sequence A of inputs $a_1, \ldots, a_n$ which is assumed to be too large to store in memory.
- We usually assume that n is not known and hence one can think of this model as a type of online or dynamic algorithm that is maintaining (say) current statistics.
- The space available $S(n)$ is some sublinear function. The input streams by and what can only store information in the sublinear space allotted.
- It is also desirable that that each input is processed efficiently, say $\log n$ and perhaps even in time $O(1)$.
- The goal is to approximately compute some function (say a statistic) of the data or identify some particular element(s) of the data stream.
- Most results concern the space required for a one pass algorithm. But there are other results concerning the tradeoff between the space and number of passes.

Some well-studied streaming problems

- Computing frequency moments. Let $A = a_1 \ldots a_n$ be a data stream with $a_i \in [m] = \{1, 2, \ldots m\}$. Let n_i denote the number of occurences of the value i in the stream A. The k^{th} frequency moment is $F_k = \sum_{i \in [m]} (n_i)^k$
  1. $F_1 = n$, the length of the sequence which can be simply computed.
  2. F_0 is the number of distinct elements in the stream
  3. F_2 is a special case of interest called the repeat index (also known as Ginis homogeneity index).
- Finding k-heavy hitters; i.e. those elements appearing at least n/k times in stream A.
- Finding rare or unique elements in A.

An example of a deterministic streaming algorithms

As in sublinear time, it will turn out that almost all of the results in this area are for randomized algorithms. Here is one exception.

The missing element problem

Suppose we are given a stream $A = a_1, \ldots, a_{n-1}$ and we are promised that the stream A is a permutation of $\{1, \ldots, n\} - \{x\}$ for some integer x in $[1, n]$. The goal is to compute the missing x.

- Space n is obvious using a bit vector $c_j = 1$ iff j has occured.
- Instead we know that $\sum_{j \in A} = n(n+1)/2 - x$.
  So if $s = \sum_{i \in A} a_i$, then $x = n(n+1)/2 - s$.
  This uses only $2 \log n$ space and constant time/item.

Generalizing to k missing elements

Now suppose we are promised a stream A of length $n-k$ whose elements consist of a permutation of $n-k$ distinct elements in $\{1,\ldots,n\}$. We want to find the missing k elements.

- Generalizing the one missing element solution, to the case that there are k missing elements we can (for example) maintain the sum of j^{th} powers $(1 \le j \le k)$ $s_j = \sum_{i \in A}(a_i)^j = c_j(n) - \sum_{i \notin A} x_i^j$. Here $c_j(n)$ is the closed form expression for $\sum_{i=1}^{n} i$. This results in k equations in k unknowns using space $k^2 \log n$ but without an efficient way to compute the solution.
- As far as I know there may not be an efficient small space streaming algorithm for this problem.
- Using randomization, much more efficient methods are known; namely, there is a streaming alg with space and time/item $O(k \log k \log n)$; it can be shown that $\Omega(k \log(n/k))$ space is necessary.

What is known about computing F_k?

Given an error bound ϵ and confidence bound δ, the goal in the frequency moment problem is to compute an estimate F'_k such that $Prob[|F_k - F'_k| > \epsilon F_k] \leq \delta$.

- The seminal paper in this regard is by Alon, Matias and Szegedy (AMS) [1999]. AMS establish a number of results:
  1. For $k \geq 3$, there is an $\tilde{O}(m^{1-1/k})$ space algorithm. (The $\tilde{O}$ notation hides factors that are polynomial in $\frac{1}{\epsilon}$ and polylogarithmic in $m, n, \frac{1}{\delta}$.
  2. For $k = 0$ and every $c > 2$, there is an $O(\log n)$ space algorithm computing F'_0 such that $Prob[(1/c)F_0 \leq F'_0 \leq cF_0$ does *not* hold$] \leq 2/c$.
  3. For $k = 1$, $\log n$ is obvious to exactly compute the length but an estimate can be obtained with space $O(\log \log n + 1/\epsilon)$
  4. For $k = 2$, they obtain space $\tilde{O}(1) = O(\frac{\log(1/\delta}{\epsilon^2}(\log n + \log m))$
  5. They also show that for all $k > 5$, there is a (space) lower bound of $\Omega(m^{1-5/k})$.

Results following AMS

- A considerable line of research followed this seminal paper. Notably settling the conjecture in AMS:
- The following results apply to real as well as integral k.
  1. An $\tilde{\Omega}(m^{1-2/k})$ space lower bound for all $k > 2$ (Bar Yossef et al [2002]).
  2. Indyk and Woodruff [2005] settle the space bound for $k > 2$ with a matching upper bound of $\tilde{O}(m^{1-2/k})$
- The basic idea behind these randomized approximation algorithms is to define a random variable Y whose expected value is close to F_k and variance is sufficiently small such that this r.v. can be calculated under the space constraint.
- We will just sketch the (non optimal) AMS results for F_k for $k > 2$ and the result for F_2.

The AMS F_k algorithm

Let $s_1 = (\frac{8}{\epsilon^2} m^{1-\frac{1}{k}})/\delta^2$ and $s_2 = 2 \log \frac{1}{\delta}$.

AMS algorithm for F_k

The output Y of the algorithm is the median of s_2 random variables $Y_1, Y_2, ..., Y_{s_2}$ where Y_i is the average of s_1 random variables $X_{ij}, 1 \leq j \leq s_1$. All X_{ij} are independent identically distributed random variables. Each $X = X_{ij}$ is calculated in the same way as follows: Choose random $p \in [1, \ldots, n]$, and then see the value of a_p. Maintain $r = |\{q | q \geq p \text{ and } a_q = a_p\}|$. Define $X = n(r^k - (r-1)^k)$.

- Note that in order to calculate X , we only require storing a_p (i.e. $\log m$ bits) and r (i.e. at most $\log n$ bits). Hence the Each $X = X_{ij}$ is calculated in the same way using only $O(\log n + \log m)$ bits.
- For simplicity we assume the input stream length n is known but it can be estimated and updated as the stream unfolds.
- We need to show that $\mathbf{E}[X] = F_k$ and that the variance is small enough to use the Chebyshev inequality.

Before moving on

- Streaming is an important field of recent algorithmic research. We have just considered one problem in one particular input model, the so-called time series model. There are two more general input models in which this and other similar problems can be studied.
  1. Cash register model: the input stream is a sequence $(I_1, I_2, , I_n)$ where $I_t = (a_j, c_t)$ and $c_t \geq 1$ representing how much to increase the (integral) count for item a_j. The current count state $(n_1(t), \ldots, n_m(t))$ at time t is then $n_i(t) = n_i(t-1) + c_t$ if $I_t = (i, c_t)$ and otherwise $n_i(t) = n_i(t-1)$.
  2. Turnstile model: this is the same model but now $|c_t| \geq 1$ allowing (integral) decrements as well as increments.

New topic: the weighted majority algorithm

I am following a survey type paper by Arora, Hazan and Kale [2005]. To quote from their paper: "We feel that this meta-algorithm and its analysis should be viewed as a basic tool taught to all algorithms students together with divide-and-conquer, dynamic programming, random sampling, and the like".

- The weighted majority algorithm and generalizations
  The "classical" WMA pertains to the following situation:
  Suppose we have say n expert weathermen (or maybe "expert" stock market forecasters) and at every time t, they give a binary prediction (rain or no rain, Raptors win or lose, dow jones up or down, Romney or Obama if there was an election every day).
- Now some or all of these experts may actually be getting their opinions from the same sources (or each other) and hence these predictions can be highly correlated.
- Without any knowledge of the subject matter (and why should I be any different from the "experts") I want to try to make predictions that will be nearly as good (over time t) as the BEST expert.

The weighted majority algorithm

The WM algorithm

Set $w_i(0) = 1$ for all i

For $t = 0...$

Our $(t+1)^{st}$ predication is

0: if $\sum_{i:\text{ expert } i \text{ predicts } 0} w_i(t) \geq (1/2)\sum_i w_i(t)$

1: if $\sum_{i:\text{ expert } i \text{ predicts } 1} w_i(t) \geq (1/2)\sum_i w_i(t)$; arbitrary o.w.

% We vote with weighted majority; arbitrary if tie

For $i = 1..n$

If expert i made a mistake on $(t+1)^{st}$ prediction

then $w_i(t+1) = (1-\epsilon)w_i(t)$;

else $w_i(t+1) = w_i(t)$

End If

End For

End For

How good is our uninformed prediction?

Theorem : Perfomance of WM

Theorem: Let $m_i(t)$ be the number of mistakes of expert i after the first t forecasts, and let $M(t)$ be the number of our mistakes. Then for any expert i (including the best expert) $M(t) \leq \frac{2 \ln n}{\epsilon} + 2(1+\epsilon)m_i(t)$.

- That is, we are essentially within a multiplicative factor of 2 plus an additive term of the best expert (without knowing anything).
- Using randomization, the factor of 2 can be removed. That is, instead of taking the weighted majority opinion, in each iteration t, choose the prediction of the i^{th} expert with probability $w_i(t)/\sum_i w_i(t)$

Theorem: Performance of Randomized WM

For any expert i, $\mathbf{E}[M(t)] \leq \frac{\ln n}{\epsilon} + (1+\epsilon)m_i(t)$

What is the meaning of the randomized impovement?

- In many applications of randomization we can argue that randomization is (provably) necessary and in other applications, it may not be provable so far but current experience argues that the best algorithm in theory and practice is randomized.
- For some algorithms (and especially online algorithms) analyzed in terms of worst case performance, there is some debate on what randomization is actually accomplishing.
- In a [1996] article Blum states that "Intuitively, the advantage of the randomized approach is that it dilutes the worst case". He continues to explain that in the determinstic algorithm, slightly more than half of the total weight could have predicted incorrectly, causing the algorithm to make a mistake and yet only reducing the total weight by $1/4$ (when $\epsilon = 1/2$). But in the randomized version, there is still a .5 probability that the algorithm will predict correctly. Convincing?

An opposing viewpoint

- In the blog LessWrong this view is strongly rejected. Here the writer makes the following comments: "We should be especially suspicious that the randomized algorithm guesses with probability proportional to the expert weight assigned. This seems strongly reminiscent of betting with 70% probability on blue, when the environment is a random mix of 70% blue and 30% red cards. We know the best bet and yet we only sometimes make this best bet, at other times betting on a condition we believe to be less probable.
  Yet we thereby prove a smaller upper bound on the expected error. Is there an algebraic error in the second proof? Are we extracting useful work from a noise source? Is our knowledge harming us so much that we can do better through ignorance?" The writer asks: "So what's the *gotcha* ... the improved upper bound proven for the randomized algorithm did not come from the randomized algorithm making systematically better predictions - doing superior cognitive work, being more intelligent - but because we arbitrarily declared that an intelligent adversary could read our mind in one case but not in the other."

Generalizing: The Multiplicative Weights algorithm

Bluem's article expresses a second benefit of the randomized approach: "Therefore the algorithm can be naturally applied when predictions are 'strategies' or other sports of things that cannot easily be combined together. Moreover, if the 'experts' are programs to be run or functions to be evaluated, then this view speeds up prediction since only one expert needs to be examined in order to produce the algorithm's prediction ...." Following the terminolgy in Arora et al, the Weighted Majority algorithm can be generalized to the multiplicative weights algorithm. We now consider that the i^{th} expert or decision on day t is a real valued cost/profit $m_i(t) \in [-1, 1]$ and let $\epsilon \leq 1/2$.

Performance of The MW algorithm

Let $\mathbf{p}(t)$ be the distribution defined by $< w_1(t), \ldots, w_n(t) >$ normalized by $\mathbf{\Phi}(t) = \sum_i w_i(t)$. Then the expected cost of the MW algorithm after T rounds is $\sum_{t=1}^{T} \mathbf{m}(t) \cdot \mathbf{p}(t) \leq \frac{\ln n}{\epsilon} + \sum_{t=1}^{T} m_i(t) + \epsilon \sum_{t=1}^{T} |m_i(t)|$

Reinterpreting in terms of gains instead of losses

We can have a vector $\mathbf{m}(t)$ of gains instead of losses and then use the "cost vector" $-\mathbf{m}(t)$ in the MW algorithm resulting in:

Performance of The MW algorithm for gains

$\sum_{t=1}^{T} \mathbf{m}(t) \cdot \mathbf{p}(t) \geq -\frac{\ln n}{\epsilon} + \sum_{t=1}^{T} m_i(t) - \epsilon \sum_{t=1}^{T} |m_i(t)|$

By taking convex combinations, an immediate corollary is

Performance wrt. a fixed distribution p

$\sum_{t=1}^{T} \mathbf{m}(t) \cdot \mathbf{p}(t) \geq -\frac{\ln n}{\epsilon} + \sum_{t=1}^{T} \mathbf{m}(t) - \epsilon |\mathbf{m}(t)|)\mathbf{p}$

An application to learning a linear binary classifier

Instead of the online application of following expert advice, let us now think of "time" as rounds in an iterative procedure. In particular, we would like to compute a linear binary classifier (when it exists).

- We are trying to classsify objects characterized by n features; that is by points $\mathbf{a}$ in $\Re^n$. We are given m labelled examples $(\mathbf{a}_1, \ell_1), \ldots, (\mathbf{a}_m, \ell_m)$ where $\ell_j \in \{-1, +1\}$
- We are going to assume that these examples can be "well classified" by a linear classifier in the sense that there exists a non negative vector $\mathbf{x}^* \in \Re^n$ (with $x_i \geq 0$) such that $sign(\mathbf{a}_j \cdot \mathbf{x}^*) = \ell_j$ for all j.
- This is equivalent to saying $\ell_j \mathbf{a}_j \cdot \mathbf{x}^* \geq 0$ and furthermore to explain the "well") we will say that $\ell_j \mathbf{a}_j \cdot \mathbf{x}^* \geq \delta$ for some $\delta > 0$.
- The goal now is to learn some linear classifer; ie a non negative $\mathbf{x} \in \Re^n$ such that $\ell_j \mathbf{a}_j \cdot \mathbf{x}^* \geq 0$. Without loss of generality, we can assume that $\sum_i x_i = 1$.
- Letting $\mathbf{b}_j = \ell_j \mathbf{a}_j$, this can now be veiwed as a reasonably general LP (search) problem.

Littlestone's Winnow algorithm for learning a linear classifier

- Litlestone [1987] used the multiplicative weights approach to solve this linear classification problem.
- Let $\rho = \max_j ||\mathbf{b}_j||_\infty$ and let $\epsilon = \delta/(2\rho)$
- The idea is to run the MW algorthm with the decisions given by the n features and gains specified by the m examples. The gain for feature i with respect to the j^{th} example is defined as $(\mathbf{b}_j)_i/\rho$ which is in [-1,1]. The $\mathbf{x}$ we are seeking is the distribution $\mathbf{p}$ in MW.

The Winnow algorithm

Initialize $\mathbf{p}$
While there are points not yet satisfied
  Let $\mathbf{b}_j \cdot \mathbf{p} < 0$ % a constraint not satisfied
  Use MW to upate $\mathbf{p}$
End While

Bound on number of iterations

The Winnow algorithm will terminate in at most $\lceil 4\rho^2 \ln n/\delta^2 \rceil$ iterations.

Miscellaneous topics to end the term

Clearly "algorithm design, analysis, and theory" is a extremely broad subject (and one might say it is much of what CS does) so clearly we have only discussed a few topic and even then only discussed them briefly. Here are a few possible topics which with we will conclude the course:

1. A return to non-oblivious local and its power.
2. The constructive Lovasz Local Lemma for 3SAT when variables do not appear in too many clauses.
3. spectral methods