CMPSCI 711: More Advanced Algorithms

Section 1-4: Sketches for ℓ_p Sampling

Andrew McGregor

Last Compiled: April 29, 2012

Fully Dynamic Vectors

- *Stream:* Consists of m updates $(x_i, \Delta_i) \in [n] \times \mathbb{R}$ that define vector f where $f_j = \sum_{i:x_i=j} \Delta_i$. E.g., for $n = 4$

$$\langle (1,3), (3,0.5), (1,2), (2,-2), (2,1), (1,-1), (4,1) \rangle$$

  defines the vector $f = (4, -1, 0.5, 1)$
- *ℓ_p Sampling:* Return random values $I \in [n]$ and $R \in \mathbb{R}$ where

$$\mathbb{P}[I = i] = (1 \pm \epsilon)\frac{|f_i|^p}{\|f\|_p^p} + n^{-c} \quad \text{and} \quad R = (1 \pm \epsilon) f_i$$

Application 1: The Social Network Puzzle

- Each person in a social network is friends with some arbitrary subset of the other $n-1$ people in the network.
- Each person only knows about their friendships.
- With no communication in the network, each person sends a postcard to Mark Zuckerberg.
- For Mark to know if the graph is connected, how long do the postcards need to be?
- We'll return to this in the next section of the course...

Application 2: Optimal F_k estimation

- Earlier we used $\tilde{O}(n^{1-1/k})$ space to (ϵ, δ) approximate $F_k = \sum_i |f_i|^k$.
- *Algorithm:* Let (I, R) be an ℓ_2 sample. Return

$$T = \hat{F}_2 R^{k-2} \qquad \text{where } \hat{F}_2 \text{ is an } e^{\pm\epsilon} \text{ estimate of } F_2$$

- *Expectation:*

$$\mathbb{E}[T] = \hat{F}_2 \sum \mathbb{P}[I = i] (e^{\pm\epsilon} f_i)^{k-2} = e^{\pm\epsilon k} F_2 \sum_{i \in [n]} \frac{f_i^2}{F_2} f_i^{k-2} = e^{\pm\epsilon k} F_k$$

- *Variance:*

$$\mathbb{V}[T] = e^{\pm 2\epsilon k} \sum \frac{f_i^2}{F_2} F_2^2 f_i^{2(k-2)} = e^{\pm 2\epsilon k} F_2 F_{2k-2} \le e^{\pm 2\epsilon k} n^{1-2/k} F_k^2$$

- *Chebychev and Chernoff:* Average $O(n^{1-2/k} \epsilon^{-2} \log \delta^{-1})$ copies.

ℓ_2 Sampling: Basic Idea

- Assume for simplicity $F_2(f) = 1$.
- Weight f_i by $\sqrt{w_i} = \sqrt{1/u_i}$ where $u_i \in_R [0,1]$ to form vector g:

$$\begin{aligned} f &= (f_1, f_2, \ldots, f_n) \\ g &= (g_1, g_2, \ldots, g_n) \quad \text{where } g_i = \sqrt{w_i} f_i \end{aligned}$$

- For some threshold t, return (i, f_i) if there is a unique i with $g_i^2 \geq t$
- Probability (i, f_i) is returned if t is sufficiently large:

$$\begin{aligned} \mathbb{P}\left[g_i^2 \geq t \text{ and } \forall j \neq i,\ g_j^2 < t\right] &= \mathbb{P}\left[g_i^2 \geq t\right] \prod_{j \neq i} \mathbb{P}\left[g_j^2 < t\right] \\ &= \mathbb{P}\left[u_i \leq \frac{f_i^2}{t}\right] \prod_{j \neq i} \mathbb{P}\left[u_j > \frac{f_j^2}{t}\right] \approx \frac{f_i^2}{t} \end{aligned}$$

- Probability some value is returned $\sum_i f_i^2/t = 1/t$ so repeating $O(t \log \delta^{-1})$ ensure a value is returned with probability $1 - \delta$.
- Unfortunately, can't store all g_i so we use Count-Sketch...

ℓ_2 Sampling: Part 1

- Use Count-Sketch with parameters (m, d) to sketch g.
- To estimate f_i^2: Let $\hat{g}_i^2 = \text{median}_j\, c_{j,h_j(i)}^2$ and $\hat{f}_i^2 = \hat{g}_i^2 / w_i$
- *Lemma:* With high probability if $d = O(\log n)$,

$$\hat{g}_i^2 = g_i^2 e^{\pm\epsilon} \pm O\left(\frac{F_2(g)}{\epsilon m}\right)$$

- *Corollary:* With high probability if $d = O(\log n)$ and $m \gg F_2(g)/\epsilon$,

$$\hat{f}_i^2 = f_i^2 e^{\pm\epsilon} \pm 1/w_i$$

- *Exercise:* $\mathbb{P}\left[F_2(g) \leq c \log n\right] \leq 99/100$ for sufficiently large $c > 0$.

Proof of Lemma

- Let $c_{j,h_j(i)} = r_j(i)g_i + Z_j$.
- By previous analysis $\mathbb{E}\left[Z_j^2\right] \leq F_2(g)/m$ and by Markov,

$$\mathbb{P}\left[Z_j^2 \leq 3F_2(g)/m\right] \geq 2/3$$

- Suppose $|g_i| \geq \frac{2}{\epsilon}|Z_j|$, then $|c_{j,h_j(i)}|^2 = e^{\pm\epsilon}|g_i|^2$
- Suppose $|g_i| \leq \frac{2}{\epsilon}|Z_j|$, then

$$|c_{j,h_j(i)}^2 - g_i^2| \leq (|g_i| + |Z_j|)^2 - |g_i|^2 = |Z_j|^2 + 2|g_i Z_j| \leq \frac{6|Z_j|^2}{\epsilon} \leq \frac{18F_2(g)}{\epsilon m}$$

  where the last inequality holds with probability $2/3$.
- Taking median over $d = O(\log n)$ repetitions, gives high probability.

ℓ_2 Sampling: Part 2

- Let $s_i = 1$ if $\hat{f_i}^2 w_i \geq 4/\epsilon$ and $s_i = 0$ otherwise
- If there is a unique i with $s_i = 1$ then return $(i, \hat{f}_i^2)$.
- Note that if $\hat{f_i}^2 w_i \geq 4/\epsilon$ then $1/w_i \leq \epsilon \hat{f}_i^2/4$ and so

$$\hat{f}_i^2 = f_i^2 e^{\pm\epsilon} \pm 1/w_i = f_i^2 e^{\pm\epsilon} \pm \epsilon \hat{f_i}^2/4$$

  and therefore $f_i^2 = e^{\pm 4\epsilon} \hat{f}_i^2$
- *Lemma:* With probability $\Omega(\epsilon)$ there's a unique i with $s_i = 1$. If there is a unique i, $\mathbb{P}[i = i^*] = e^{\pm 8\epsilon} f_{i^*}^2$.
- *Thm:* Repeat $O(\epsilon^{-1} \log n)$ times. Total space: $O(\epsilon^{-2} \operatorname{polylog} n)$.

Proof of Lemma

- Let $t = 4/\epsilon$. We can upper-bound $\mathbb{P}[s_i = 1]$:

$$\mathbb{P}[s_i = 1] = \mathbb{P}\left[\hat{f}_i^2 w_i \geq t\right] \leq \mathbb{P}\left[e^{4\epsilon} f_i^2/t \geq u_i\right] \leq e^{4\epsilon} f_i^2/t$$

  and similarly, $\mathbb{P}[s_i = 1] \geq e^{-4\epsilon} f_i^2/t$.
- Assuming independence of w_i, probability of unique i with $s_i = 1$:

$$\begin{aligned}
\sum_i \mathbb{P}\left[s_i = 1, \sum_{j \neq i} s_j = 0\right] &\geq \sum_i \mathbb{P}[s_i = 1]\,(1 - \sum_{j \neq i} \mathbb{P}[s_j = 1]) \\
&\geq \sum_i \frac{e^{-4\epsilon} f_i^2}{t}\left(1 - \frac{\sum_{j \neq i} e^{4\epsilon} f_i^2}{t}\right) \\
&\geq \frac{e^{-4\epsilon}(1 - e^{4\epsilon}/t)}{t} \approx 1/t
\end{aligned}$$

- If there is a unique i, probability $i = i^*$ is

$$\frac{\mathbb{P}\left[s_{i^*} = 1, \sum_{j \neq i^*} s_j = 0\right]}{\sum_i \mathbb{P}\left[s_i = 1, \sum_{j \neq i} s_j = 0\right]} = e^{\pm 8\epsilon} f_{i^*}^2$$

ℓ_0-Sampling

- Maintain $\tilde{F}_0$, an $(1 \pm .1)$-approximation to F_0.
- Hash items using $h_j : [n] \to [0, 2^j - 1]$ for $j \in [\log n]$.
- For each j, maintain:

$$D_j = (1 \pm 0.1)|\{t|h_j(t) = 0\}|$$

$$S_j = \sum_{t, h_j(t)=0} f_t i_t$$

$$C_j = \sum_{t, h_j(t)=0} f_t$$

- *Lemma:* At level $j = 2 + \left\lceil \log \tilde{F}_0 \right\rceil$ there is an *unique* element in the stream that maps to 0 with constant probability.
- Uniqueness is verified if $D_j = 1 \pm 0.1$. If unique, then S_j / C_j gives identity of the unique element and C_j is the count.

Proof of Lemma

- Let $j = \left\lceil \log \tilde{F}_0 \right\rceil$ and observe that $2F_0 < 2^j < 12F_0$.
- For any i, $\mathbb{P}[h_j(i) = 0] = 1/2^j$.
- Probability there exists a unique i such that $h_j(i) = 0$,

$$
\begin{aligned}
& \sum_i \mathbb{P}[h_j(i) = 0 \text{ and } \forall k \neq i,\ h_j(k) \neq 0] \\
= \quad & \sum_i \mathbb{P}[h_j(i) = 0] \, \mathbb{P}[\forall k \neq i,\ h_j(k) \neq 0 | h_j(i) = 0] \\
\geq \quad & \sum_i \mathbb{P}[h_j(i) = 0] \, (1 - \sum_{k \neq i} \mathbb{P}[h_j(k) = 0 | h_j(i) = 0]) \\
= \quad & \sum_i \mathbb{P}[h_j(i) = 0] \, (1 - \sum_{k \neq i} \mathbb{P}[h_j(k) = 0]) \\
\geq \quad & \sum_i \frac{1}{2^j}(1 - \frac{F_0}{2^j}) \geq \frac{1}{24}
\end{aligned}
$$

- Note that the above holds true even if h_j is only 2-wise independent.