# CMPSCI 711: More Advanced Algorithms 

Vectors 4: Sketches for $\ell_{p}$ Sampling

Andrew McGregor

Last Compiled: February 16, 2018

## Fully Dynamic Vectors

- Stream: Consists of $m$ updates $\left(x_{i}, \Delta_{i}\right) \in[n] \times \mathbb{R}$ 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 $f=(4,-1,0.5,1)$

- $\ell_{p}$ Sampling: Return random values $I \in[n]$ and $R \in \mathbb{R}$ where

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

## 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 $\tilde{O}\left(n^{1-1 / k}\right)$ space to $(\epsilon, \delta)$ approximate $F_{k}=\sum_{i}\left|f_{i}\right|^{k}$.
- Algorithm: Let $(I, R)$ be an $\ell_{2}$ sample. Return

$$
T=\hat{F}_{2} R^{k-2} \quad \text { where } \hat{F}_{2} \text { is an } e^{ \pm \epsilon} \text { estimate of } F_{2}
$$

- Expectation:

$$
\mathbb{E}[T]=\hat{F}_{2} \sum \mathbb{P}[I=i]\left(e^{ \pm \epsilon} f_{i}\right)^{k-2}=e^{ \pm \epsilon k} F_{2} \sum_{i \in[n]} \frac{f_{i}^{2}}{F_{2}} f_{i}^{k-2}=e^{ \pm \epsilon k} F_{k}
$$

- Variance:

$$
\mathbb{V}[T]=e^{ \pm 2 \epsilon k} \sum \frac{f_{i}^{2}}{F_{2}} F_{2}^{2} f_{i}^{2(k-2)}=e^{ \pm 2 \epsilon k} F_{2} F_{2 k-2} \leq e^{ \pm 2 \epsilon k} n^{1-2 / k} F_{k}^{2}
$$

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


## $\ell_{2}$ 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 $g$ :

$$
\begin{aligned}
f & =\left(f_{1}, f_{2}, \ldots, f_{n}\right) \\
g & =\left(g_{1}, g_{2}, \ldots, g_{n}\right) \quad \text { where } g_{i}=\sqrt{w_{i}} f_{i}
\end{aligned}
$$

- For some threshold $t$, return $\left(i, f_{i}\right)$ if there is a unique $i$ with $g_{i}^{2} \geq t$
- Probability $\left(i, f_{i}\right)$ 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 $O\left(t \log \delta^{-1}\right)$ ensure a value is returned with probability $1-\delta$.
- Unfortunately, can't store all $g_{i}$ so we use Count-Sketch...


## $\ell_{2}$ Sampling: Part 1

- Use Count-Sketch with parameters ( $m, d$ ) to sketch $g$.
- To estimate $f_{i}^{2}$ : Let $\hat{g}_{i}^{2}=\operatorname{median}_{j} c_{j, h_{j}(i)}^{2}$ 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,

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

- Suppose $\left|g_{i}\right| \geq \frac{2}{\epsilon}\left|Z_{j}\right|$, then $\left|c_{j, h_{j}(i)}\right|^{2}=e^{ \pm \epsilon}\left|g_{i}\right|^{2}$
- Suppose $\left|g_{i}\right| \leq \frac{2}{\epsilon}\left|Z_{j}\right|$, then

$$
\left|c_{j, h_{j}(i)}^{2}-g_{i}^{2}\right| \leq\left(\left|g_{i}\right|+\left|Z_{j}\right|\right)^{2}-\left|g_{i}\right|^{2}=\left|Z_{j}\right|^{2}+2\left|g_{i} Z_{j}\right| \leq \frac{6\left|Z_{j}\right|^{2}}{\epsilon} \leq \frac{18 F_{2}(g)}{\epsilon m}
$$

where the last inequality holds with probability $2 / 3$.

- Taking median over $d=O(\log n)$ repetitions, gives high probability.


## $\ell_{2}$ 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 $\left(i, \hat{f}_{i}^{2}\right)$.
- 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 $i, \mathbb{P}\left[i=i^{*}\right]=e^{ \pm 8 \epsilon} f_{i^{*}}^{2}$.
- Thm: Repeat $O\left(\epsilon^{-1} \log n\right)$ times. Total space: $O\left(\epsilon^{-2}\right.$ polylog $\left.n\right)$.


## Proof of Lemma

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

$$
\mathbb{P}\left[s_{i}=1\right]=\mathbb{P}\left[\hat{f}_{i}^{2} w_{i} \geq t\right] \leq \mathbb{P}\left[e^{4 \epsilon} f_{i}^{2} / t \geq u_{i}\right] \leq e^{4 \epsilon} f_{i}^{2} / t
$$

and similarly, $\mathbb{P}\left[s_{i}=1\right] \geq e^{-4 \epsilon} f_{i}^{2} / t$.

- Assuming independence of $w_{i}$, probability of unique $i$ with $s_{i}=1$ :

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

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

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

## $\ell_{0}$-Sampling

- Maintain $\tilde{F}_{0}$, an ( $1 \pm .1$ )-approximation to $F_{0}$.
- Hash items using $h_{j}:[n] \rightarrow\left[0,2^{j}-1\right]$ for $j \in[\log n]$.
- For each $j$, maintain:

$$
\begin{gathered}
D_{j}=(1 \pm 0.1)\left|\left\{t \mid h_{j}(t)=0\right\}\right| \\
S_{j}=\sum_{t, h_{j}(t)=0} f_{t} i_{t} \\
C_{j}=\sum_{t, h_{j}(t)=0} f_{t}
\end{gathered}
$$

- 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 $2 F_{0}<2^{j}<12 F_{0}$.
- For any $i, \mathbb{P}\left[h_{j}(i)=0\right]=1 / 2^{j}$.
- Probability there exists a unique $i$ such that $h_{j}(i)=0$,

$$
\begin{aligned}
& \sum_{i} \mathbb{P}\left[h_{j}(i)=0 \text { and } \forall k \neq i, h_{j}(k) \neq 0\right] \\
= & \sum_{i} \mathbb{P}\left[h_{j}(i)=0\right] \mathbb{P}\left[\forall k \neq i, h_{j}(k) \neq 0 \mid h_{j}(i)=0\right] \\
\geq & \sum_{i} \mathbb{P}\left[h_{j}(i)=0\right]\left(1-\sum_{k \neq i} \mathbb{P}\left[h_{j}(k)=0 \mid h_{j}(i)=0\right]\right) \\
= & \sum_{i} \mathbb{P}\left[h_{j}(i)=0\right]\left(1-\sum_{k \neq i} \mathbb{P}\left[h_{j}(k)=0\right]\right) \\
\geq & \sum_{i} \frac{1}{2^{j}}\left(1-\frac{F_{0}}{2^{j}}\right) \geq \frac{1}{24}
\end{aligned}
$$

- Note that the above holds true even if $h_{j}$ is only 2 -wise independent.

