Sketching Algorithms for Big Data Fall 2017

Lecture 8 — September 26, 2017

Prof. Jelani Nelson *Scribe: Sebastian Gehrmann*

1 Overview

1. Continuous monitoring
2. Chaining

2 Continuous Monitoring

In previous lectures, we considered problems in which we observe a stream of data and then query information about it. In continuous monitoring (CM), we consider data streams with continuous, more involved querying. An an example, consider a router that sees a stream of IP's and at every point in time we are interested in what the heavy hitters are, how skewed the distribution is, or whether we can detect trends and anomalies in the traffic. In addition, we are interested in identifying when the answer to a query changes.

Dynamic Data Structures Refresher Generally, data structures (DS) are ways of laying out data in memory. A dynamic DS is a structure that allows updates, opposed to a static structure that does not allow them. A stream is an example of a dynamic DS (with the goal of sublinear memory).

Problem Description formalize the CM problem, we consider an operation sequence that looks like $up_1, q_1, up_2, q_2, \ldots, up_n, q_n$ for updates up_i and queries q_i. So far, we have considered Monte-Carlo randomized algorithms (that fail with a certain probability). We typically want $f(x)$ and algorithmic output $\tilde{f}$ s.t. $\mathbb{P}(|f(x) - \tilde{f}| > \gamma) < \delta$ (*).

In CM, we have $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, where $x^{(t)}$ is the frequency vector after the first t updates. An algorithm should output $\tilde{f}^{(1)}, \tilde{f}^{(2)}, \ldots, \tilde{f}^{(m)}$ s.t. $\mathbb{P}(\exists_{t\in[m]} |\tilde{f}^{(t)} - f(x^{(t)})| > \gamma_{(t)}) < \delta$ (**). We can derive (**) from (*) by setting $\delta' = \frac{\delta}{m}$, and then take the union bound oder all $t \in [m]$

However, by doing this, we blow up the space and query time by $\log n$. The goal is to prevent this from happening?

Approach Our hope is to define an event E_t that algorithmically fails at time t ($\tilde{f}$ is a bad estimate). We want to bound

$$\mathbb{P}(\bigvee_{t=1}^{m} E_t) = \sum_{t=1}^{m} \mathbb{P}(E_t | \bar{E}_1 \wedge \bar{E}_2 \wedge \ldots \wedge \bar{E}_{t-1})$$

So far, most algorithms we have seen are unlikely to fail at time $t+1$ if they worked t times. An example for this is the sketch by Alon, Matias, and Szegedy [1] for insertion-only streams (each update does $c \leftarrow x+1$ for some i). Let us assume that we only have one row. Then, we maintain a random sign vector σ (as opposed to a matrix) that we maintain the dot-product with stream X with. The intuition is now that the dot-product does not change much over time.

$$\underbrace{\mathbb{E}|X| \leq (\mathbb{E}X^2)^{\frac{1}{2}}}_{\text{Jensen's inequality}} = (\mathbb{E}(\sum_{i=1}^{t} \sigma_i)^2)^{\frac{1}{2}} = \sqrt{t}$$

The trick here is to see that $\mathbb{E}(\sum_{i=1}^{t} \sigma_i)^2$ is 0 for off-diagonals while the diagonals are ± 1, which means that the expectation is t.

Toy Example - Random Walk on a Line

For demonstration purposes, we are going to assume the following problem. We have a line (. . . , -4, -3, -2, -1, 0, 1, 2, 3, 4, . . .). At each step we either take a step right or left With equal probability. We run this for N time steps. The evolution can then be described by a row of the AMS sketch. If we define

$$x^{(t)} = (\overbrace{1, 1, \ldots, 1}^{t}, \overbrace{0, \ldots, 0}^{N-t})$$

then at time t, we are at position $\langle \sigma, x^{(t)} \rangle$, $\sigma \in -1, 1^N$, and $\mathbb{E}|\langle \sigma, x^{(N)} \rangle| \leq \sqrt{N}$ (same as above). We want to guarantee that we never leave this $\sqrt{N}$ area:

$$\mathbb{E} \sup_{t \in [N]} |\langle \sigma, x^{(N)} \rangle| = O(\sqrt{N})$$

We have $\mathbb{P}^{(t)} = \frac{x^{(t)}}{\sqrt{N}}$, and want $\mathbb{E} \sup_t |\langle \sigma, x \rangle| = O(1)$

We are going to prove this via chaining (which is not the usual way). For the proof we need the Khintchine inequality:

$$\forall x \in \mathbb{R}^N \ \mathbb{P}(\langle \sigma, x \rangle > \lambda) < \exp(\frac{-\lambda^2}{\delta \|x\|_2^2})$$

We can restate the dot product as $s \sum_i \sigma_i x_i$. Now, using Markov:

$$\mathbb{P}(e^{s \sum_i \sigma_i x_i} > e^{s\lambda}) < e^{s\lambda} \mathbb{E} e^{s \sum_i \sigma_i x_i}$$

We can now calculate an upper bound on the last expectation in this equality:

$$
\begin{aligned}
\mathbb{E}e^{s\sum_i \sigma_i x_i} &= \mathbb{E}\prod_i e^{s\sum_i \sigma_i x_i} \\
&= \prod_i \mathbb{E}e^{s\sum_i \sigma_i x_i} && \text{Linearity of expectations} \\
&= \prod_i \frac{1}{2}(e^{sx_i} + e^{-sx_i}) && \text{Taylor exp. of } e = \sum_{k=1}^{\infty} \frac{z^k}{k} \\
&= \prod_i \cosh(sx_i) \\
&\leq \prod_i e^{\frac{s^2 x_i^2}{2}} \\
&= e^{\frac{s^2\|x\|_2^2}{2-s\lambda}} && \text{choose } s = \frac{\lambda}{\|x\|_2^2}
\end{aligned}
$$

3 Chaining

Chaining is a method for computing bounds on $\mathbb{E}\sup_{z\in Z} Z$ that leverages correlations across the Z. For the lecture today, every Z looks like $\langle\sigma, x\rangle$ for some x. We want to bound

$$\boxed{\mathbb{E}\sup_{x\in T}\langle\sigma, x\rangle \stackrel{\text{def}}{=} r(T)}$$

The reasoning is that if you bound this and the negative of this, we have a guarantee. If you have two vectors x and y where $|x - y|$ is close to 0, then $\langle\sigma, x\rangle$ is close to $\langle\sigma, y\rangle$.

3.1 Bounding $r(t)$

We are going to explore four ways of bounding $r(T)$ that are increasingly tight. We will use the property that $r(T) \lesssim \text{rad}(T) * \lg^{\frac{1}{2}} |T|$.

(1) - Union Bound First, we normalize T so that all $x \in T$ have $\|x\|_2^2 \leq 1$ (we can later reverse this). We can then use the fact that $\mathbb{E}\, Z \leq \mathbb{E}\,|Z| \leq \int_0^\infty \mathbb{P}(|Z| > u)du$

$$r(T) \leq \int_0^\infty \mathbb{P}(\sup_{x\in T}\langle\sigma, x\rangle > u)du$$

We can then break up the integral and use the fact that the full integral over a distribution is 1.

$$\underbrace{\int_0^{C\sqrt{\lg|T|}} \mathbb{P}(\ldots)du}_{\leq C\sqrt{\lg|T|}} + \int_{C\sqrt{\lg|T|}}^\infty \mathbb{P}(\exists x \in T|\langle\sigma, x\rangle| > u)du$$

Since the left part is bounded, we can focus on the right part and union bound the tail down.

$$\begin{aligned}\int_{C\sqrt{\lg|T|}}^{\infty}\mathbb{P}(\exists x\in T|\langle\sigma,x\rangle|>u)du &\leq \int_{C\sqrt{\lg|T|}}^{\infty}\sum_{x\in T}\mathbb{P}(|\langle\sigma,x\rangle|>u)du\\ &\leq 2|T|\int_{C\sqrt{\lg|T|}}^{\infty}e^{\frac{-u^2}{2}}du\\ &= O(1)\end{aligned}$$

We have found bounds for both terms, but we do not achieve the desired overall $O(1)$ due to the left term.

(2) - ϵ-net argument Let T' be the smallest ϵ-net of T under l_2. For $x\in T$, let $x'\in T'$ be the closest point to x in T'. Then

$$\begin{aligned}\mathbb{E}\sup_{x\in T}\langle\sigma,x\rangle &= \mathbb{E}\sup_{x\in T}\langle\sigma,x'+(x-x')\rangle\\ &\leq r(T')+\mathbb{E}\sup_{x\in T}\langle\sigma,x-x'\rangle\end{aligned}$$

Here, the first term is $\leq \lg^{\frac{1}{2}}\mathcal{N}(T,l_2,\epsilon)$. Using Cauchy-Schwarz and property of the ϵ-net, we can bound the second term to $\leq \|\sigma\|_2\|x-x'\|_2\leq\epsilon\sqrt{N}$. Now we choose the best ϵ:

$$r(T)\leq\inf_{\epsilon\to 0}\{\lg^{\frac{1}{2}}\mathcal{N}(T,l_2,\epsilon)+\epsilon\sqrt{N}\}$$

Remember that

$$x^{(t)}=\overbrace{(1,1,\ldots,1}^{t},\overbrace{0,\ldots,0)}^{N-t}$$

Now we need to estimate $\mathcal{N}(T,l_2,\epsilon)$. We plug in the representation above to receive

$$(0,\frac{(1,0,\ldots,0)}{\sqrt{N}},\frac{(1,0,\ldots,0)}{\sqrt{N}},\ldots,\frac{(1,1,\ldots,1)}{\sqrt{N}})$$

For two time-points $t>s$, $r^{(t)}-r^{(s)}=\sqrt{\frac{t-s}{N}}$, since

$$\|r^{(t)}-r^{(s)}\|_2=\|\frac{\overbrace{(0,0,1,1,\ldots,0,1)}^{t-s}}{\sqrt{N}}\|_2=\sqrt{\frac{t-s}{N}}$$

This implies that $\mathcal{N}(T,l_2,\epsilon)\leq\frac{1}{\epsilon^2}$. If we plug this in above, we arrive at the same guarantee as in the union-bound approach.

(3) - Dudley's inequality For this approach, we assume that $\text{rad}(T) \leq 1$. The idea is that instead of stopping after the expansion $x' + (x - x')$, we recurse and create more ϵ-nets. Let S_k be a $\frac{1}{2^k}$-net of T under l_2 (the choice does not matter for this example, but it does for others). Let $x^{(k)}$ be the closest point to x in S_k. Now the full expansion becomes

$$\langle \sigma, x \rangle = \langle \sigma, x^{(0)} + \sum_{k=1}^{\infty} x^{(k)} - x^{(k-1)} \rangle$$

Now we can use Dudley's inequality [2] to compute an upper bound.

$$\begin{aligned}
\mathbb{E} \sup_{x \in T} \langle \sigma, x \rangle &= \mathbb{E} \sup_{x \in T} \langle \sigma, x^{(0)} + \sum_{k=1}^{\infty} x^{(k)} - x^{(k-1)} \rangle \\
&\leq \underbrace{\mathbb{E} \langle \sigma, 0 \rangle}_{=0} + \underbrace{\mathbb{E} \sup_{x \in T} \sum_{k=1}^{\infty} \langle \sigma, x^{(k)} - x^{(k-1)} \rangle}_{\text{switching makes worse}} \\
&\leq \sum_{k=1}^{\infty} \mathbb{E} \sup_{x \in T} \underbrace{\langle \sigma, x^{(k)} - x^{(k-1)} \rangle}_{\text{norm} \leq \frac{3}{2^k} = \frac{1}{2^k} + \frac{1}{2^{k-1}}} \\
&\lesssim \sum_{k=1}^{\infty} \frac{1}{2^k} \lg^{\frac{1}{2}} (\mathcal{N}(T, l_2, \frac{1}{2^k}) \cdot \mathcal{N}(T, l_2, \frac{1}{2^{k-1}})) \\
&\lesssim \sum_{k=1}^{\infty} \frac{1}{2^k} \lg^{\frac{1}{2}} \mathcal{N}(T, l_2, \frac{1}{2^k}) \ \leftarrow (*) \text{ Dudley's inequality} \\
&\left(\leq \int_0^{\infty} \lg^{\frac{1}{2}} \mathcal{N}(T, l_2, u) du \right)
\end{aligned}$$

RWL: $(*) \lesssim \sum_{k=1}^{\infty} \frac{\sqrt{k}}{2^k} = O(1)$

This shows that the random walk on a line is bounded.

(4) - Last approach (not full proof) First, observe the following. Say $T_0 \leq T_1 \leq T_2 \leq \ldots \leq T$ is "admissible" if (a) $|T_0| = 1$, (b) $|T_r| \leq 2^{2^r}$. The claim now is that

$$(*) \simeq \inf_{\{T_r\}_{r=0}^{\infty}} \sum_{r=1}^{\infty} \underbrace{2^{\frac{r}{2}}}_{\sqrt{\log} \text{ of net in (3)}} \sup_{x \in T} d_{l_2}(x, T_r)$$

The intuition for the sup is that it stays close to 1, then drops to $\leq .5$. Altogether it gives you the same bound as (3). The idea behind this proof is to find the best possible solution with a budget of 2^{2^r}.

Generic Chaining The technique of generic chaining can be traced back to Fernique [3]. The idea is required to get from the random walk on a line example to streaming

$$r(T) \lesssim \inf_{\{T_r\}} \sup_{x \in T} \sum_{r=1}^{\infty} 2^{\frac{r}{2}} d_{l_2}(x, T_r)$$

Now, $x^{(t)}$ refers to where the stream is after the first t updates. We claim that if we define $r^{(t)} = \frac{x^{(t)}}{\|x^{(m)}\|_2}$, then

$$\forall u \in (0,1)\ \mathcal{N}(\{r^{(t)}\}_{t=1}^{m}, l_2, u) \leq \frac{1}{u^2}$$

The proof for this is similar to the line using chaining and Dudley's inequality. We can use this to estimate l_2 in a stream. Let the output at time t be $\tilde{f}^{(t)}$.

Definition 1. *We achieve weak tracking if* $\forall t \in [m] |\tilde{f}^{(t)} - \|x^{(t)}\|_2| < \epsilon \|x^{(m)}\|_2$

Definition 2. *We achieve strong tracking if* $\forall t \in [m] |\tilde{f}^{(t)} - \|x^{(t)}\|_2| < \epsilon \|x^{(t)}\|_2$

We can use today's techniques to show that the AMS sketch with $k = O(\frac{1}{\epsilon^2 \lg \frac{1}{\epsilon}}$ rows provides weak tracking (and it can be improved to $k = O(\frac{1}{\epsilon^2}$ with a different chaining argument by Braverman et al. [4]).

Why is this relevant? AMS maintains $y_j = \langle \sigma(j), x \rangle$ for $j = 1, \ldots, k$. $\|x\|_2^2$ is estimated as $\sum_j y_j^2 = \sum_j \langle \sigma(j), x \rangle^2$. We can use the media of $\lg \frac{1}{d}$ repetitions to get a low failure probability. We can argue that $\langle \sigma(j), x^{(t)} - x^{(s)} \rangle$ is never big when we plug it into the equation from the line-walking example:

$$\begin{aligned} \langle \sigma(j), x^{(t_2)} \rangle &= \langle \sigma(j), x^{(t_1)} + (x^{(t_2)} - x^{(t_1)}) \rangle \\ &= \langle \sigma(j), x^{(t_1)} \rangle + \langle \sigma(j), x^{(t_2)} - x^{(t_1)} \rangle \end{aligned}$$

References

[1] Noga Alon, Yossi Matias, Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[2] Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. of Func. Anal.*, 1(3):290–330.

[3] Xavier Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. *Ecole dEté de Probabilités de Saint-Flour*, 1–96, 1975.

[4] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, David P. Woodruff. BPTree: An l_2 Heavy Hitters Algorithm Using Constant Memory. *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 361–376, 2017.