10 Mar 2025

Quantile estimation, Reservoir Sampling, Glivenko–Cantelli

If tokens in a stream come from an ordered set, e.g. $[m]$, we can define the quantile function a.k.a. the <u>empirical CDF</u>

$$q(a) = \frac{\#\{i \mid a_i \le a\}}{n}$$

\# stream tokens $\le a$

total length of stream.

Two goals for quantile sketching.
The algorithm estimates $q(a)$ using estimator $\hat{q}(a)$.

<u>(ε,δ)-PAC</u>: $\forall a \; \Pr\left(|\hat{q}(a) - q(a)| > \varepsilon\right) < \delta$.

<u>Uniformly (ε,δ)-PAC</u>:

$$\Pr\left(\|\hat{q} - q\|_\infty > \varepsilon\right) < \delta$$

where $\|\hat{q} - q\|_\infty$ denotes $\max_{a \in [m]} |\hat{q}(a) - q(a)|$.

This lecture in a nutshell:

1. <u>Downsampling</u> the stream — maintaining in memory a random sample of $t \approx \frac{1}{\varepsilon^2}$ elements — is a good way to maintain a quantile sketch.

2. <u>Reservoir sampling</u> achieves downsampling in $O(t + \log_m(n))$ space.

3. Given $t \approx \frac{1}{\varepsilon^2}$ uniformly random elements of the stream, the empirical CDF of those t elements is an ε-approx to the empirical CDF of the entire stream.

Glivenko–Cantelli Theorem

DKW Inequality

Dvoretzky – Kiefer – Wolfowitz

Reservoir Sampling for $t=1$ random elements

Store a_1 in memory

for $s = 2, 3, \ldots, n$:

with probability $\frac{1}{s}$ overwrite stored elt with a_s

Probability that a_i is stored at the end?

- pick a_i and store it when we first see it. $\left(\frac{1}{i}\right)$

- don't pick a_s for all $s > i$.

$$\left(1-\frac{1}{i+1}\right)\cdot\left(1-\frac{1}{i+2}\right)\cdot \cdots \cdot \left(1-\frac{1}{n}\right).$$

Product of all these probabilities:

$$\frac{1}{\cancel{i}}\cdot\left(\frac{\cancel{i}}{\cancel{i+1}}\right)\cdot\left(\frac{\cancel{i+1}}{\cancel{i+2}}\right)\cancel{}\cdots\left(\frac{\cancel{n-1}}{n}\right)=\frac{1}{n}.$$

Generalization to selecting t out of n.

1. Initialize empty buffer of size t.
   $b = (b_1, \ldots, b_t) = (\perp, \perp, \ldots, \perp)$.

2. for $s = 1, \ldots, n$:
   if $s \le t$ store $b_s = a_s$
   if $s > t$:
   with probability $\frac{t}{s}$:
   sample $i \in [t]$ uniformly random
   overwrite $b_i \leftarrow a_s$
   else: // probability $1-\frac{t}{s}$
   do nothing

Fact. If $n \geq t$ then the buffer contents after processing stream of length n

are a uniformly random t-element subset.

Quantile estimation using reservoir sampling.

Maintain reservoir sample $b_1, \dots, b_t$ in memory.

When queried about $g(a)$ for $a \in [m]$ report

$$\hat{g}(a) = \frac{\#\{i \mid b_i \le a\}}{t}$$

a.k.a. $\hat{g}$ is the empirical CDF of the reservoir sample.

Analyzing the error $|\hat{g}(a) - g(a)|$

We'll analyze an algorithm where $b_1, \dots, b_t$ are independent, each is uniformly distributed.

Space: $O(t + \log_m(n))$

storing $b_1, \dots, b_t$

storing the counter $s \in [n]$ when one element of $[m]$ takes $O(1)$ storage.

$$\Pr\left(|\hat{g}(a) - g(a)| > \varepsilon\right) < 2e^{-2\varepsilon^2 t}$$

by Hoeffding Bound.

$$X_i = \begin{cases} 1 & \text{if } b_i \le a \\ 0 & \text{o.w.} \end{cases} \qquad i \in [t]$$

b_i is unif random

$$\mathbb{E}[X_i] = \Pr(b_i \le a) = \frac{\#\{s \mid a_s \le a\}}{n} = q(a)$$

Let $X = X_1 + \cdots + X_t = t \cdot \hat{q}(a)$

the event $|\hat{q}(a) - q(a)| > \varepsilon$ is equiv't to

$$|X - \mathbb{E}X| > \varepsilon t.$$

$b_1, \ldots, b_t$ indep't $\Longrightarrow$ $X_1, \ldots, X_t$ indep't

$$\xrightarrow{\text{(Hoeffding)}} \Pr(|X - \mathbb{E}X| > \varepsilon t) < 2e^{-2\varepsilon^2 t^2 / t}$$

<u>Summary</u>: For (ε, δ)-PAC quantile est
maintain t indep. unif. random samples
from the stream where t is chosen
s.t.

$$2e^{-2\varepsilon^2 t} < \delta$$

$$t > \frac{1}{2}\varepsilon^{-2} \ln\left(\frac{2}{\delta}\right)$$

Upgrading to a uniform (ε,δ)-PAC estimate

Keep same algorithm. Vary the value of t according to the analysis technique.

Plan 1. Union bound

There are m events of the form

$$|\hat{q}(a) - q(a)| > \varepsilon$$

one for each $a \in [m]$.

We can make each have probability less than $\frac{\delta}{m}$ by setting

$$t > \frac{1}{2}\varepsilon^{-2}\ln\left(\frac{2m}{\delta}\right).$$

Plan 2. More careful union bound

Lemma. (weak DKW) If $b_1, \dots, b_t$ are independent ident distrib samples from any distribution on $\mathbb{R}$ with CDF $F(a)$ and if we define the emprical CDF

$$\hat{q}(a) = \frac{\#\{i \mid b_i \le a\}}{t}$$

then

$$\Pr\left(\|\hat{q} - F\|_\infty > \varepsilon\right) < \frac{4}{\varepsilon} e^{-\varepsilon^2 t/2}.$$

Compare with

$$\Pr\left(\|\hat{g}-F\|_\infty > \varepsilon\right) < 2m\, e^{-2\varepsilon^2 t}$$

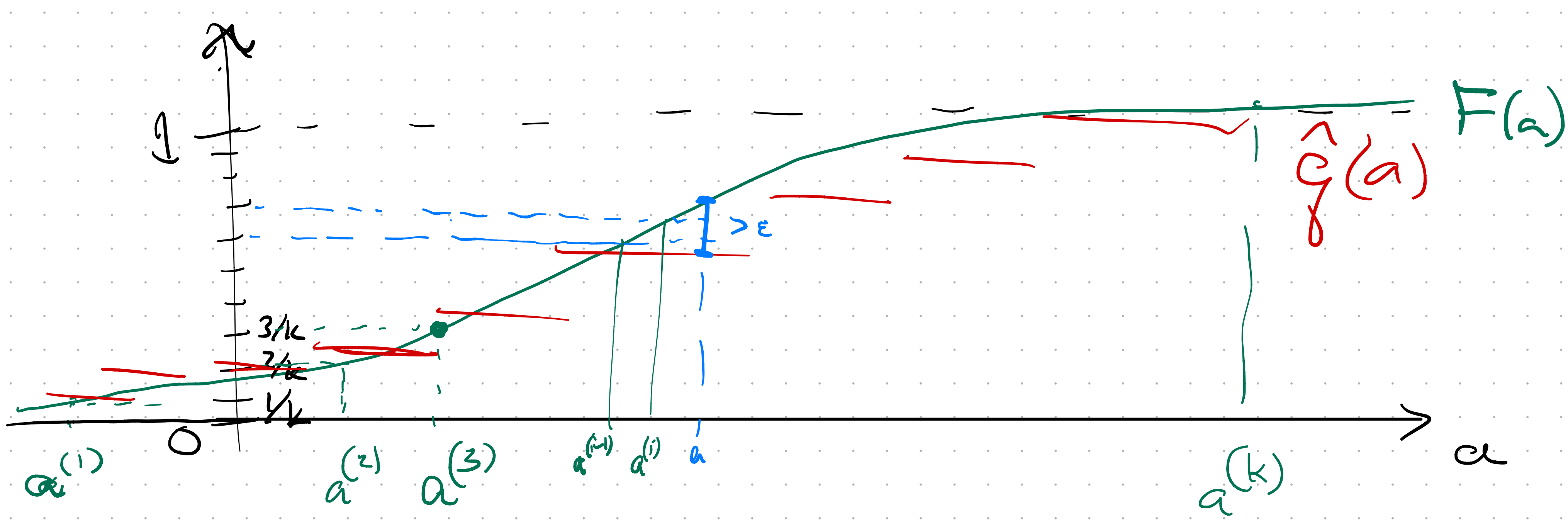
where F supported on $[m]$ and we use naive union bound a la Plan 1.

Proof of weak DKW.

Let $k = \left\lceil \frac{2}{\varepsilon} \right\rceil$.

We'll find a set of $2(k-1) < \frac{4}{\varepsilon}$ bad events, such that if none of them happen,

$$\|\hat{g} - F\|_\infty \leq \varepsilon.$$



$$a^{(i)} = \sup\left\{ a \,\middle|\, F(a) < \frac{i}{k} \right\}.$$

Proof shows that if $\exists a$ s.t.

$$|\hat{g}(a) - F(a)| > \varepsilon$$ then

$\exists i \in [k-1]$ s.t.

$$F(a^{(i)}) - \hat{g}(a^{(i)}) > \varepsilon/2 \quad \text{or}$$

$$\check{g}(a^{(i)}) - F(a^{(i)}-) > \varepsilon/2$$

$$\check{g}(a^{(i)}) = \sup\{\hat{g}(a) \mid a < a^{(i)}\} \qquad F(a^{(i)}-) = \sup\{F(a) \mid a < a^{(i)}\}$$