9 Mar 2026

Reservoir Sampling
Quantile Estimation
Glivenko – Cantelli Theorem

Announcements. (1) Prob Set 1 & Quiz 4 grades released
Regrades available for 2 weeks (PSet)
1 week (quiz)

(2) Typos corrected on PSet 2.
See Ed post.

Puzzle. You're going to see a stream of n tokens, each in $\{0,1\}^b$.
You only have space to store one token and one counter in range $0, 1, \ldots, n$.
How to output a uniformly random element of the token sequence?

Idea. Store one token, y, in memory.
When processing x_t, replace y with x_t with "some probability such that it remains uniformly random."

RESERVOIR SAMPLING:

initialize $y = x_1$.

for t in $2, 3, 4, \ldots, n$:

observe token x_t.

with probability $1/t$:

$y \leftarrow x_t$

else:

y remains unchanged

endfor

output y.

Claim $\forall t \geq 1$, after t rounds of this loop, y is unif distrib over $\{x_1, \ldots, x_t\}$.

Proof Induct on t.

Base case $t=1$: $y = x_1$ ✓

Induct step: after loop iteration $t+1$,

$$\Pr(y = x_{t+1}) = \frac{1}{t+1} \text{ by design.}$$

For any $1 \leq s \leq t$,

$$\Pr(y = x_s) = \Pr(y \text{ unchanged in iteration } t+1) \cdot \Pr(y = x_s \text{ at start of iter})$$

$$= \left(1 - \frac{1}{t+1}\right) \cdot \frac{1}{t}$$

$$= \frac{1}{t+1}.$$

RESERVOIR SAMPLING for drawing a k-element subset of $\{x_1, \ldots, x_n\}$ uniformly at random from all k-element subsets.

initialize $(y_1, \ldots, y_k) = (x_1, \ldots, x_k)$

for $t = k+1, k+2, \ldots, n$:

with probability $\frac{k}{t}$:

// store x_t in $\vec{y}$, overwriting a random stored element

Sample index $i \in [k]$ unif. random

$y_i \leftarrow x_t$

else:

ignore x_t, leave $\vec{y}$ unchanged.

Correctness proof: again by induction.

(omitted, but in book.)

Space requirement: $k \cdot b$ + $\log m$

$y_1, \ldots, y_k$ t.

Quantile estimation.

Tokens are elements of $[N] = \{1, 2, \ldots, 2^b\}$.

Goal. Store a sketch of stream $x_1, \ldots, x_m$ and afterward answer quantile queries:

$$q(i): \frac{\#\{j \mid x_j \le i\}}{m}$$

(ε, δ)-PAC quantile est.:

On query i, output $\hat{q}(i)$ s.t.

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

(ε, δ)-PAC uniform quantile est:

Output a data structure encoding a vector of quantile estimates

$(\hat{q}(0), \hat{q}(1), \ldots, \hat{q}(m))$ such that

$$\Pr\left(\exists i \text{ with } |\hat{q}(i) - q(i)| > \varepsilon\right) < \delta.$$

Clearly (ε,δ)-PAC UQE $\Rightarrow$ (ε,δ)-PAC QE

$(\varepsilon, \frac{\delta}{N})$-PAC QE $\Rightarrow$ (ε,δ)-PAC UQE.

First let's talk about QE.

Claim: For appropriate choice of k, the following algo works:

- use reservoir sampl to downsample $x_1, \ldots, x_m$ to $y_1, \ldots, y_k$.
- using $y_1, \ldots, y_k$ simulate k indep. stream samples with replacement $z_1, \ldots, z_k$
- answer query $q(i)$ with

$$\hat{q}(i) = \frac{\#\{\ell \mid z_\ell \le i\}}{k}$$

How large must k be for this to work?

To show. $|\hat{q}(i) - q(i)|$ likely small.

Idea. Use Hoeffding.

$$X_\ell = \begin{cases} 1 & \text{if } \underbrace{y_\ell}_{z_\ell} \le i \\ 0 & \text{otherwise} \end{cases}$$

$q(i)$.

Indeed $E(X_\ell) = \frac{1}{n} \cdot \#\{j \mid x_j \le i\}$

because y_ℓ is unif distn

over $\{x_1, \dots, x_n\}$.

$$\hat{q}(i) = \frac{1}{k}(X_1 + X_2 + \dots + X_k).$$

$$\Pr\left(|\hat{q}(i) - q(i)| > \varepsilon \right)$$

$$= \Pr\left(|k\cdot\hat{q}(i) - k\cdot q(i)| > k\varepsilon \right)$$

$$= \Pr\left(|(X_1 + \dots + X_k) - E[X_1 + \dots + X_k]| > k\varepsilon \right)$$

(Hoeff)

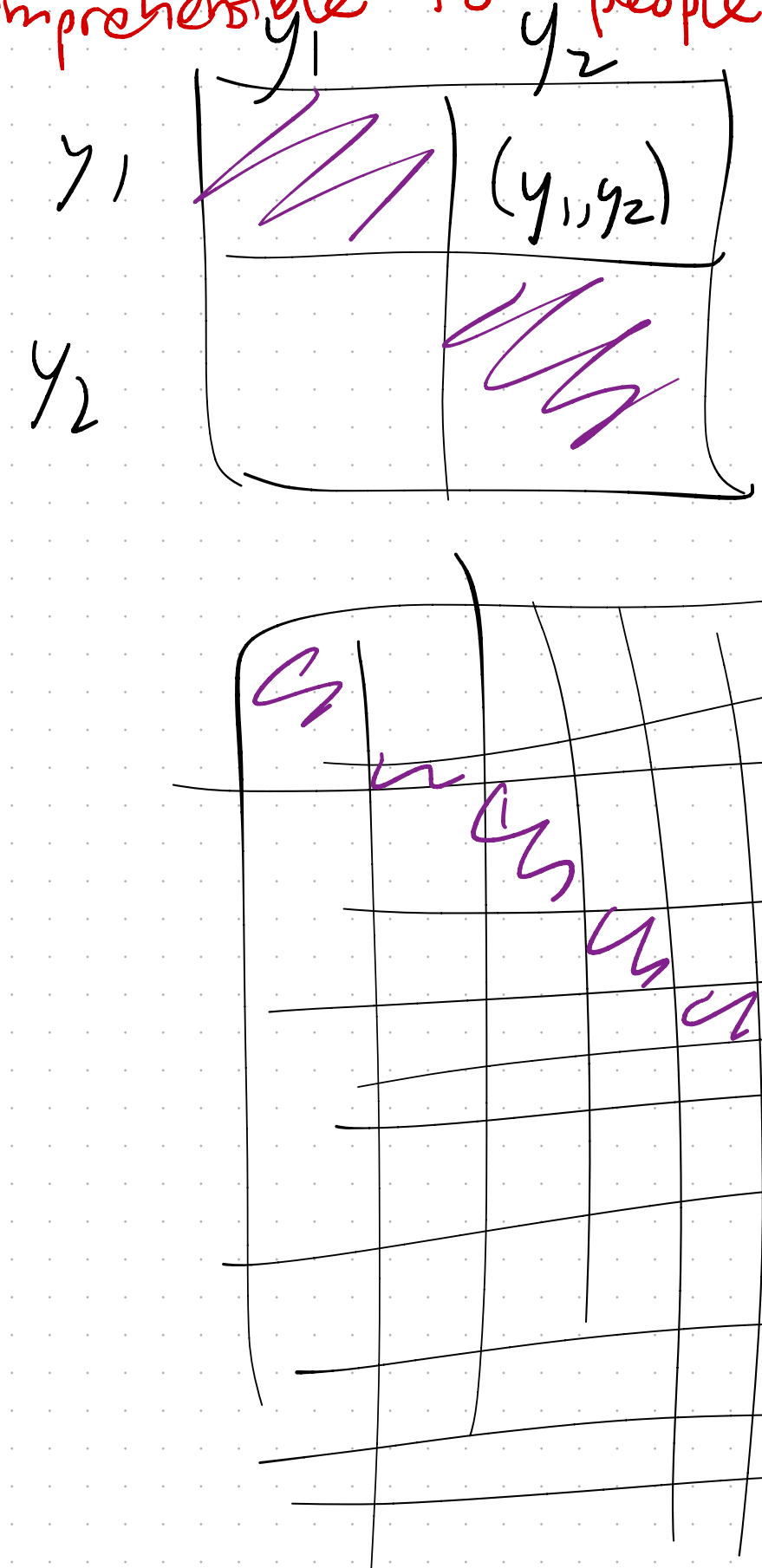
$$\le 2e^{-\frac{2(k\varepsilon)^2}{k}} = 2e^{-2k\varepsilon^2}$$

To make this $< \delta$, set

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

<BEGIN "SCRAP PAPER" DIGRESSION>

Readers of these notes after the lecture should ignore this part of the notes. They were used as a visual aid but aren't meant to be comprehensible to people who weren't present.



$k=2$:

$$z_1 = \begin{cases} y_1 & \text{w. prob } 1/2 \\ y_2 & \text{--- } 1/2 \end{cases}$$

$$z_2 = \begin{cases} z_1 & \text{w. prob. } \frac{1}{m} \\ \{y_1, y_2\} - \{z_1\} & \text{w. prob. } \frac{m-1}{m}. \end{cases}$$

More generally.

After sampling $z_1, \ldots, z_{l-1}$;

let d = # distinct values in $(z_1, \ldots, z_{l-1})$.

$$z_l = \begin{cases} \text{w prob } \frac{d}{m} & \text{repeat } \overset{\text{random}}{\wedge}\text{one of the } d \text{ distinct vals} \\ \text{w. prob. } \frac{m-d}{m} & \text{simple } \overset{\text{random}}{\wedge}\text{one of the unsampled elements of tuple } (y_1, \ldots, y_m). \end{cases}$$

{END SCRAP PAPER DIGRESSION}

Better still.

Run k indep copies of reservoir sampling (each sampling step stream element) to yield $(z_1, \ldots, z_k)$ each iid uniform over $\{x_1, \ldots, x_m\}$.

Then $\hat{q}(i) = \frac{\#\{l \mid z_l \le i\}}{k}$.

Summary. sample size

$$k = \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

suffices for (ε, δ)-PAC QE,

$$\Rightarrow k = \frac{1}{2\varepsilon^2} \ln\left(\frac{2N}{\delta}\right)$$

suffices for (ε, δ)-PAC UQE.

In fact

$$k = \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

samples are sufficient even for (ε, δ)-PAC UQE.

DKW inequality (Dvoretzky–Kiefer–Wolfowitz)

If $z_1, \ldots, z_k$ are iid draws from a distribution with CDF F, and $\hat{F}_z$ denotes the empirical CDF

$$\hat{F}_z(a) = \frac{\#\{l : z_l \le a\}}{k}$$

then $$\Pr\left(\|\hat{F}_z - F\|_\infty > \varepsilon \right) < 2e^{-2\varepsilon^2 k}$$

N.B. For a single $a \in \mathbb{R}$,

$$\Pr\left(|\hat{F}_z(a) - F(a)| > \varepsilon \right) < 2e^{-2\varepsilon^2 k}.$$

(Like Hoeffding bd + union bd combined without requiring union bd.)

$\Longrightarrow$ (ε,δ)-PAC UQE requires space

$$O\left(\log m + \frac{1}{\varepsilon^2} \ln\left(\frac{1}{\delta}\right) b \right)$$

⟹ Glivenko–Cantelli Theorem

If $Z_1, Z_2, \ldots$ is an infinite sequence of iid samples from distn. with CDF F and

$$\hat{F}_t(a) = \#\{s \mid 1 \le s \le t,\ Z_s \le a\} / t$$

= empirical CDF of first t samples

then $\Pr\left(\|\hat{F}_t - F\|_\infty \to 0 \text{ as } t \to \infty \right) = 1$

Reason. DKW implies $\forall \varepsilon > 0$

$$\mathbb{E}\left[\#t : \|\hat{F}_t - F\|_\infty > \varepsilon \right] \text{ is finite.}$$

$\Longrightarrow$ w. prob. 1,

$\|\hat{F}_t - F\|_\infty \le \varepsilon$ for all suff large t.