■ **Exercise 1 (A streaming algorithm for the second moment of the frequencies).** We are given a stream of numbers $x_1, \ldots, x_n \in \{0, \ldots, m-1\}$ and we want to compute the sum of the squares of the frequencies of each values 0 to $m-1$ in this stream: if $f_a(x) = \#\{i : x_i = a\}$, we want to compute $F_2(x) = \sum_{a=0}^{m-1} f_a(x)^2$.

Take a random function $h : \{1, \ldots, m\} \to \{-1, 1\}$, i.e.: for all a, $h(a)$ is chosen at independently and uniformly at random in $\{-1, 1\}$. And do the following:

Algorithm 1 Second frequency moment random algorithm

Pick a hash function $h : \{0, \ldots, m-1\} \to \{-1, 1\}$ uniformly at random
Compute $Z = h(x_1) + \cdots + h(x_n)$ while reading the stream
Output Z^2

▶ ***Question 1.1)*** *Show that $\mathbb{E}[Z^2] = F_2(x)$ where the expectation is taken over all the possible values for h.* ▷ Hint. *For $a \neq b$, show that $\mathbb{E}_h[h(a)h(b)] = 0$ and $\mathbb{E}_h[h(a)^2] = 1$.*

Answer. ▷ First remark that, $\mathbb{E}[h(a)^2] = \mathbb{E}[1] = 1$ and $\mathbb{E}[h(a)] = \frac{1}{2} \times -1 + \frac{1}{2} \times 1 = 0$. Now, if $a \neq b$, then $h(a)$ and $h(b)$ are independent and $\mathbb{E}[h(a)h(b)] = \mathbb{E}[h(a)]\,\mathbb{E}[h(b)] = 0$. It follows:

$$\begin{aligned}
\mathbb{E}[Z^2] = \mathbb{E}[(h(x_1) + \cdots h(x_n))^2] &= \mathbb{E}\left[\left(\sum_{a=0}^{m-1} f_a(x)h(a)\right)^2\right] \\
&= \sum_{a=0}^{m-1} f_a(x)^2\,\mathbb{E}[h(a)^2] + \sum_{a \neq b} f_a(x)f_b(x)\,\mathbb{E}[h(a)h(b)] \\
&= \sum_{a=0}^{m-1} f_a(x)^2 = F_2(x).
\end{aligned}$$

◁

▶ ***Question 1.2)*** *Show that $\mathbb{V}ar(Z^2) = \mathbb{E}[Z^4] - \mathbb{E}[Z^2]^2 = 2\sum_{a \neq b} f_a(x)^2 f_b(x)^2 \leqslant 2F_2(x)^2$.*

Answer. ▷ As before, note that if b, c, d are all different from a, by independence of $h(a)$ from $h(b)$, $h(c)$ and $h(d)$, we have: $\mathbb{E}[h(a)^3 h(b)] = \mathbb{E}[h(a)h(b)] = 0$ and

$\mathbb{E}[h(a)h(b)h(c)h(d)] = \mathbb{E}[h(a)]\,\mathbb{E}[h(b)h(c)h(d)] = 0$. It follows that:

$$\begin{aligned}
\mathbb{E}[Z^4] &= \mathbb{E}\left[\left(\sum_{a=0}^{m-1} f_a(x)h(a)\right)^4\right] \\
&= \sum_{a=0}^{m-1} f_a(x)^4\,\mathbb{E}[h(a)^4] \\
&\quad + 4\sum_{a=0}^{m-1}\sum_{b\neq a} f_a(x)^3 f_b(x)\,\mathbb{E}[h(a)^3 h(b)] \\
&\quad + \frac{1}{2}\binom{4}{2}\sum_{a=0}^{m-1}\sum_{b\neq a} f_a(x)^2 f_b(x)^2\,\mathbb{E}[h(a)^2 h(b)^2] \\
&\quad + \binom{4}{2}\sum_{a=0}^{m-1}\sum_{\text{distinct } b,c\neq a} f_a(x)^2 f_b(x) f_c(x)\,\mathbb{E}[h(a)^2 h(b)h(c)] \\
&\quad + \sum_{a=0}^{m-1}\sum_{\text{distinct } b,c,d\neq a} f_a(x)f_b(x)f_c(x)f_d(x)\,\mathbb{E}[h(a)h(b)h(c)f(d)] \\
&= \sum_{a=0}^{m-1} f_a(x)^4 + 3\sum_{a,b:a\neq b} f_a(x)^2 f_b(x)^2.
\end{aligned}$$

Thus, $\mathbb{V}ar[Z^2] = \mathbb{E}[Z^4] - \mathbb{E}[Z^2] = \sum_{a=0}^{m-1} f_a(x)^4 + 3\sum_{a,b:a\neq b} f_a(x)^2 f_b(x)^2 - \left(\sum_{a=0}^{m-1} f_a(x)^2\right)^2$

$$= 2\sum_{a,b:a\neq b} f_a(x)^2 f_b(x)^2 \leqslant 2\left(\sum_{a=0}^{m-1} f_a(x)^2\right)^2 = 2F_2(x)^2.$$

◁

Remark that this algorithm requires a lot of memory to store h: $O(m \log m)$ bits, almost as much as counting the frequencies independently ($O(m \log n)$ bits). But remark that we only need the values of h to be 4-*wise* independent to obtain the results above. Let us thus use the following construction for h that will require much less memory.

Consider the field $\mathbb{F}_{2^k}$ where $k = \lceil \log_2 m \rceil$ such that $2^{k-1} < m \leqslant 2^k$. Let us identify the elements of $\mathbb{F}_{2^k}$ as a string of k bits and as numbers from 0 to $2^k - 1$ as well. Let $\pi : \mathbb{F}_{2^k} \to \{-1, 1\}$ be the function that associates to any number $a \in \mathbb{F}_{2^k}$ the value -1 if the first bit of a is 0 and the value $+1$ otherwise.

For all 4-tuple $(u, v, w, t) \in (\mathbb{F}_{2^k})^4$, let $P_{uvwt} : \mathbb{F}_{2^k} \to \mathbb{F}_{2^k}$ be the polynomial:

$$P_{uvwt}(a) = ua^3 + va^2 + wa + t,$$

and set $h_{uvwt}(a) = \pi(P_{uvwt}(a))$.

► ***Question* 1.3)** *Show that if u, v, w, t are chosen independently and uniformly at random in $\mathbb{F}_{2^k}$, then for all fixed* distinct *values $a, b, c, d \in \mathbb{F}_{2^k}$, the random 4-tuple $(P_{uvwt}(a), P_{uvwt}(b), P_{uvwt}(c), P_{uvwt}(d))$ is uniform in $(\mathbb{F}_{2^k})^4$.*

Answer. ▷ Let $(p,q,r,s) \in (F_{2^k})^4$.

$$\Pr_{u,v,w,t}\{(P_{uvwt}(a),P_{uvwt}(b),P_{uvwt}(c),P_{uvwt}(d)) = (p,q,r,s)\}$$
$$= \Pr_{u,v,w,t}\left\{\begin{pmatrix} a^3 & a^2 & a & 1\\ b^3 & b^2 & b & 1\\ c^3 & c^2 & c & 1\\ d^3 & d^2 & d & 1\end{pmatrix}\begin{pmatrix} u\\ v\\ w\\ t\end{pmatrix} = \begin{pmatrix} p\\ q\\ r\\ s\end{pmatrix}\right\}$$
$$= \frac{\#\left\{(u,v,w,t) \in (F_{2^k})^4 : \begin{pmatrix} a^3 & a^2 & a & 1\\ b^3 & b^2 & b & 1\\ c^3 & c^2 & c & 1\\ d^3 & d^2 & d & 1\end{pmatrix}\begin{pmatrix} u\\ v\\ w\\ t\end{pmatrix} = \begin{pmatrix} p\\ q\\ r\\ s\end{pmatrix}\right\}}{\#(F_{2^k})^4}$$
$$= \frac{1}{(2^k)^4},$$

indeed, the solution (u,v,w,t) is unique since the matrix is a van der Mond matrix which is inversible as soon as a, b, c and d distinct. It follows that all the values in $(\mathbb{F}_{2^k})^4$ are equally probable for the 4-tuple $(P_{uvwt}(a),P_{uvwt}(b),P_{uvwt}(c),P_{uvwt}(d))$, it is thus uniform. ◁

► ***Question 1.4)*** *Conclude that when u,v,w,t are chosen independently and uniformly at random in $\mathbb{F}_{2^k}$, the values $h_{uvwt}(0),\ldots,h_{uvwt}(m-1)$ are 4-wise independent uniform random variables with values in $\{-1,1\}$.*

Answer. ▷ Remark that π maps half the elements in $\mathbb{F}_{2^k}$ to -1 and the other half to 1. Thus, the image by π of a uniform random variable in $\mathbb{F}_{2^k}$ is a uniform random variable in $\{-1,1\}$. Formally, for all $(\alpha,\beta,\gamma,\delta) \in \{-1,1\}^4$ and distincts $a,b,c,d \in \mathbb{F}_{2^k}$,

$$\Pr_{u,v,w,t}\{(\pi(P_{uvwt}(a)),\pi(P_{uvwt}(b)),\pi(P_{uvwt}(c)),\pi(P_{uvwt}(d))) = (\alpha,\beta,\gamma,\delta)\}$$
$$= \sum_{\substack{(p,q,r,s)\,\in\,\mathbb{F}_{2^k}\\ (\pi(p),\pi(q),\pi(r),\pi(s)) = (\alpha,\beta,\gamma,\delta)}} \Pr_{u,v,w,t}\{(P_{uvwt}(a),P_{uvwt}(b),P_{uvwt}(c),P_{uvwt}(d)) = (p,q,r,s)\}$$
$$= (2^{k-1})^4 \cdot \frac{1}{(2^k)^4} = \frac{1}{2^4},$$

which implies that $\pi(P_{uvwt}(0)),\ldots,\pi(P_{uvwt}(m-1))$ are 4-wise independent uniform random variables in $\{-1,1\}$. ◁

► ***Question 1.5)*** *Conclude that there is a (ε,δ)-estimator computing $F_2(x)$ using $O(\log m + \log n)$ bits of memory. Describe it and explain the bound on the memory needed as a function of δ and ε.*

Answer. ▷ Consider the following algorithm and let us prove that it is a (ε,δ)-estimator: Recall that $\mathbb{V}ar(\mu_i) = \mathbb{V}ar(Z)/B \leqslant 2F_2(x)^2/B$. By Chebychev inequality, for all $i = 1..A$,

$$\Pr\{|\mu_i - F_2(x)| \geqslant \varepsilon F_2(x)\} \leqslant \frac{\mathbb{V}ar(\mu_i)}{\varepsilon^2 F_2(x)^2} \leqslant \frac{2}{B\varepsilon^2} \leqslant \frac{1}{4}.$$

Furthermore, if the median of the values $\mu_1,\ldots,\mu_A$ lies outside $(1\pm\varepsilon)F_2(x)$, then at least $A/2$ of the values lie outside as well. Then, if Y_i denotes the indicator random variable for the event $\mu_i \notin (1\pm\varepsilon)F_2(x)$ (note that $\mathbb{E}[Y_i] \leqslant 1/4$), then by Hoeffding inequality,

$$\begin{aligned}\Pr\{|output - F_2(x)| \geqslant \varepsilon F_2(x)\} &\leqslant \Pr\{Y_1 + \cdots + Y_A \geqslant A/2\}\\ &\leqslant \Pr\{Y_1 + \cdots + Y_A - \mathbb{E}[Y_1 + \cdots + Y_A] \geqslant A/4\}\\ &\leqslant \exp\left(-\frac{2(A/4)^2}{A}\right) \leqslant \delta.\end{aligned}$$

Algorithm 2 Memory efficient second frequency moment (ε, δ)-estimator

Let $k = \lceil \log_2 m \rceil$, $A = \lceil 8 \ln(1/\delta) \rceil$ and $B = \lceil 8/\varepsilon^2 \rceil$
for $i = 1..A$ and $j = 1..B$ **do**
    Pick $u_{ij}, v_{ij}, w_{ij}, t_{ij}$ independently and uniformly at random in the field $\mathbb{F}_{2^k}$
    Let h_{ij} be the hash function: $h_{ij}(a) = \pi(u_{ij}a^3 + v_{ij}a^2 + w_{ij}a + t_{ij})$
Compute $Z_{ij} = h_{ij}(x_1) + \cdots + h_{ij}(x_n)$ for all $i = 1..A$ and $j = 1..B$ simultaneously while reading the stream
for $i = 1..A$ **do**
    Compute the average $\mu_i = \dfrac{(Z_{i1})^2 + \cdots + (Z_{iB})^2}{B}$
return the median of the values $\mu_1, \ldots, \mu_A$

Now, the algorithm is memory efficient since it uses: $4AB$ variables of k bits each (the $u_{ij}, v_{ij}, w_{ij}, t_{ij}$) and $AB + A$ variables of at most $2 \log n$ bits (the Z_{ij} and μ_i). The total number of bits of memory used by the (ε, δ)-estimator for $F_2(x)$ is thus: $O\left(\frac{\ln(1/\delta)}{\varepsilon^2}(\log m + \log n)\right)$. ◁