■ **Exercise 1 (Missing element & distinct elements).** Assume we are reading a stream of n distinct integers in $\{1, \dots, n+1\}$.

▶ ***Question* 1.1)** *Assume first that all of the elements in the stream are indeed distinct elements of $\{1, \dots, n+1\}$) and design for this case a deterministic $O(\log n)$ bits-memory algorithm that outputs the missing element.*
Answer. ▷ Just compute $S = \sum_{i=1}^{n} x_i$ and output $\frac{(n+1)(n+2)}{2} - S$. This requires at most $\lceil \log_2 \frac{(n+1)(n+2)}{2} \rceil \leqslant \lceil 2 \log_2 n \rceil$ bits of memory to store S. ◁

Let us now waive the assumption that the integers are distinct and let us design an algorithm to check this property.

▶ ***Question* 1.2)** *Consider a prime number $p \geqslant n^2$ and a non-zero polynomial $U(X)$ of degree at most n over the field $\mathbb{Z}_p$. Show that $\mathsf{Pr}_a\{U(a) = 0 \bmod p\} \leqslant \frac{1}{n}$ when a is chosen uniformly at random in $\mathbb{Z}_p$.*
▷ Hint. *How many solutions are there to $U(a) = 0$ in the field $\mathbb{Z}_p$?*

Answer. ▷ As $\mathbb{Z}_p$ is a field, a non-zero polynomial of degree d has at most d roots. It follows that $U(a) = 0$ admits at most n solutions. Thus, $\mathsf{Pr}_a\{U(a) = 0 \bmod p\} \leqslant \frac{n}{p} \leqslant \frac{1}{n}$. ◁

Consider the following algorithm: Pick a prime number p such that $n^2 \leqslant p < 2n^2$ (there is always one). Pick an integer $a \in \{0, \dots, p-1\}$ uniformly at random. Compute $S := \sum_{i=1}^{n} x_i$, $y := \frac{(n+1)(n+2)}{2} - S$, $U := \sum_{i=1}^{n} a^{x_i - 1} \bmod p$ and $V := \sum_{i=0}^{n} a^i \bmod p$. If $U == V - a^{y-1} \bmod p$, then answer « *y is the missing element* », and answer « *the stream does not contain n distinct integers in $\{1, \dots, n+1\}$* » otherwise.

▶ ***Question* 1.3)** *Show that this is a $O(\log n)$ bits-memory streaming algorithm that always outputs the right answer when the stream matches the specification, and that detects every erroneous stream with probability at least $1 - 1/n$.*
Answer. ▷ Assume that all the element in the stream are distinct integers in $\{1, \dots, n+1\}$, then by the previous question, y is indeed the missing element and the difference between U and V is indeed a^{y-1}.

Assume now that the elements in the stream are not all distinct. Then, the difference of the polynomials $U(X) = \sum_{i=1}^{n} X^{x_i - 1}$ and $V_y(x) = \sum_{i=0}^{n} X^i - X^{y-1}$ is a non-zero polynomial in $\mathbb{Z}_p$ whatever y is in $\mathbb{Z}_p$. It follows that $U(a) = U \neq V - a^{y-1} = V_y(a)$ with probability at least $1 - 1/n$ by the previous question. ◁

■ **Exercise 2 (Traffic monitoring: uniformity detection).** Imagine that we are running a huge website and we want to prevent attacks by keeping track of the origins of the various clients currently connected to the server. Along time, clients connect and then disconnect from the website. And we want to detect if all the clients connected are from the same IP address. But we do not want to slow down the server and wish to dedicate to this task only a *constant* memory, i.e. only a constant number of integers. We model the problem as follows:

We are given an infinite stream of events $e_1, e_2, \dots, e_n, \dots$ where each e_i is either **connect**(x) or **disconnect**(x) where x is a positive integer standing for the IP address of the client (dis-)connecting. We assume that the stream is wellformed, i.e. that there are always at least as many events **connect**(x) as **disconnect**(x) from the beginning of the stream to any position for every integer x. We want to detect when all the clients connected have the same IP address x.

▶ ***Question* 2.1)** *Spot when to set the alarm on in the following sequence where x denotes the event* **connect**(x) *and $\bar{x}$ the event* **disconnect**(x)*:*

$$1, 2, 3, \bar{2}, \bar{3}, 1, 1, \bar{1}, 4, 6, 7, \bar{1}, \bar{6}, \bar{1}, 2, \bar{2}, \bar{4}, 8, 3, \bar{3}, \bar{7}, 9$$

Answer. ▷ The sets of currently connected clients are (an * spots every date when all the clients connected have the same IP address):

$1:1^*$	$\bar{3}:1^*$	$4:114$	$\bar{6}:147$	$\bar{4}:7^*$	$\bar{7}:8^*$
$2:12$	$1:11^*$	$6:1146$	$\bar{1}:47$	$8:78$	$9:89$
$3:123$	$1:111^*$	$7:11467$	$2:247$	$3:378$	
$\bar{2}:13$	$\bar{1}:11^*$	$\bar{1}:1467$	$\bar{2}:47$	$\bar{3}:78$	

◁

We consider the following algorithm that uses only three integer variables:

- start with $\mathtt{n} := 0$, $\mathtt{a} := 0$ and $\mathtt{b} := 0$ at $t = 0$;
- on event **connect**(x): do $\mathtt{n} := \mathtt{n} + 1$, $\mathtt{a} := \mathtt{a} + x$ and $\mathtt{b} := \mathtt{b} + x^2$;
- on event **disconnect**(x): do $\mathtt{n} := \mathtt{n} - 1$, $\mathtt{a} := \mathtt{a} - x$ and $\mathtt{b} := \mathtt{b} - x^2$;
- set on the alarm every time that $\mathtt{n} > 0$ and $\mathtt{b} = \mathtt{a}^2/\mathtt{n}$.

The right way to the correctness of this deterministic algorithm passes through the analysis of a random variable. Consider a random variable X taking positive integer values. We denote by $\mathsf{supp}(X) = \{\, x : \mathsf{Pr}\{X = x\} > 0 \,\}$ and assume that $|\mathsf{supp}(X)| < \infty$. We denote by $\mathbb{E}[X]$ and $\mathbb{V}\mathsf{ar}[X] = \mathbb{E}[(X - \mathbb{E}(X))^2]$ respectively the expectation and the variance of X.

► ***Question* 2.2)** *Show that $|supp(X)| = 1$ if and only if $\mathbb{V}\mathsf{ar}[X] = 0$.*
Answer. ▷ First remark that for all integer valued random variable X, $supp(X) \neq \varnothing$. If $supp(X) = \{x\}$, then $\mathsf{Pr}\{X = x\} = 1$ and $\mathbb{E}[X] = x$ and $\mathbb{V}\mathsf{ar}[X] = 0$. Assume now that $|\,supp(X)| \geqslant 2$, there are $x, x' \in supp(X)$ such that $x \neq x'$, and thus either $\mathbb{E}[X] \neq x$ or $\mathbb{E}[X] \neq x'$ (or both). Assume that $\mathbb{E}[X] \neq x$. The random variable $Z = (X - \mathbb{E}[X])^2$ only takes non-negative values. Thus, $\mathbb{V}\mathsf{ar}[X] = \mathbb{E}[Z] = \sum_y \mathsf{Pr}\{X = y\} \cdot (y - \mathbb{E}[X])^2 \geqslant \underbrace{\mathsf{Pr}\{X = x\}}_{>0} \cdot \underbrace{(x - \mathbb{E}[X])^2}_{>0} > 0$. ◁

► ***Question* 2.3)** *Conclude that the algorithm is correct.*
Answer. ▷ Let us fix some time t, and let T denote the multiset of the IP adresses of the people currently connected to the server at time t. Assume that $|T| \geqslant 1$. We want to decide if T contains only the same integer. let X be the uniform random variable over the multiset T. Since $supp(X) = T$, by the previous question, $\mathbb{V}\mathsf{ar}(X) = 0$ if and only if T contains only the same integer. But, at time t, $\mathtt{n} = |T|$ and $\mathbb{E}[X] = \frac{1}{\mathtt{n}} \sum_{x \in T} x = \mathtt{a}/\mathtt{n}$ and $\mathbb{V}\mathsf{ar}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\big[\, X^2 - 2\,\mathbb{E}[X]X + \mathbb{E}[X]^2 \,\big] = \mathbb{E}[X^2] - \mathbb{E}[X] = \frac{1}{\mathtt{n}} \sum_{x \in T} x^2 - \mathtt{a}^2/\mathtt{n}^2 = \mathtt{b}/\mathtt{n} - \mathtt{a}^2/\mathtt{n}^2 = 0$ if and only if $\mathtt{b} = \mathtt{a}^2/\mathtt{n}$. The algorithm detects thus correctly when all the clients connected have the same IP address. ◁

■ **Exercise 3 ((ε, δ)-estimator).** Suppose we want to compute a value μ from some data. Assume that we have a randomized algorithm A that computes a random variable Z such that $\mathbb{E}[Z] = \mu$ and $\mathbb{V}\mathsf{ar}(Z) \leqslant A \cdot \mu^2$ for some constant $A > 0$.

► ***Question* 3.1)** *Design a (ε, δ)-estimator for μ for all $\varepsilon > 0$ and $\delta > 0$ making $O(\frac{\log(1/\delta)}{\varepsilon^2})$ calls to the randomized algorithm A. Give exact bounds on the number of calls, explain how you proceed.*
Answer. ▷ We proceed as usual by outputting the median Y of k averages of ℓ independent runs Z_{ij} of A for $i \in [k]$ and $j \in [\ell]$. Let us denote by $\mu_i = \frac{Z_{i1} + \cdots + Z_{i\ell}}{\ell}$. $\mathbb{E}[\mu_i] = \mathbb{E}[Z] = \mu$ and $\mathbb{V}\mathsf{ar}(\mu_i) = \frac{\mathbb{V}\mathsf{ar}(Z_{i1}) + \cdots + \mathbb{V}\mathsf{ar}(Z_{i\ell})}{\ell^2} = \frac{\mathbb{V}\mathsf{ar}(Z)}{\ell} \leqslant \frac{A\mu^2}{\ell}$. By Chebychev inequality, $\mathsf{Pr}\{|\mu_i - \mu| \geqslant \varepsilon\mu\} \leqslant \frac{\mathbb{V}\mathsf{ar}(\mu_i)}{\varepsilon^2\mu^2} \leqslant \frac{A}{\ell\varepsilon^2} \leqslant \frac{1}{4}$ as soon as $\ell \geqslant \frac{4A}{\varepsilon^2}$.

Now, let X_i be the indicator random variable for the event $\mu_i \notin (1 \pm \varepsilon)\mu$. Then, $\mathbb{E}[X_i] = \mathsf{Pr}\{|\mu_i - \mu| \geqslant \varepsilon\mu\} \leqslant \frac{1}{4}$. Not that if the median $Y \notin (1 \pm \varepsilon)\mu$ then at least $\frac{k}{2}$

variables among $\mu_1, \ldots, \mu_k$ do not belong to $(1 \pm \varepsilon)\mu$, i.e. $X_1 + \cdots + X_k \geqslant \frac{k}{2}$. It follows that:

$$\begin{aligned}
\Pr\{|Y-\mu| \geqslant \varepsilon\mu\} &\leqslant \Pr\{X_1 + \cdots + X_k \geqslant \frac{k}{2}\} \\
&\leqslant \Pr\left\{X_1 + \cdots + X_k - \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_k] \geqslant \frac{k}{4}\right\} \text{ since } \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_k] \leqslant \frac{k}{4} \\
&\leqslant \exp\left(-\frac{2(k/4)^2}{k}\right) \text{ by Hoeffding's inequality} \\
&\leqslant \delta,
\end{aligned}$$

as soon as $k \geqslant 8\ln(1/\delta)$. Thus taking $\ell = \lceil 4A/\varepsilon^2 \rceil$ and $k = \lceil 8\ln(1/\delta) \rceil$ yields the desired (ε, δ)-estimator for μ. ◁

Assume now that we have an algorithm B that computes a random variable Y such that $\Pr\{Y \notin [\frac{\mu}{a}, a \cdot \mu]\} \leqslant b$ for some $a > 1$ and $b < \frac{1}{2}$. (No assumption is made on $\mathbb{E}[Y]$, it might even be $\neq \mu$)

▶ ***Question* 3.2)** *Can we design a (ε, δ)-estimator for μ using algorithm B for all $\varepsilon > 0$ and $\delta > 0$? If not, what are the values of ε and δ for which we can design a (ε, δ)-estimator and how do you proceed?*

Answer. ▷ As we have no information on the expected value nor the variance of Y, it is impossible to improve on the precision ε and the best ε we can get is $\varepsilon = \max(a - 1, 1 - 1/a)$. Now, as $b < \frac{1}{2}$, we can use k runs of Y and output their median W. For the same reason as before, using Hoeffding's inequality, one can show that the probability that $W \notin [\mu/a, a \cdot \mu]$ is at most $\exp(-2(k/2 - bk)^2/k) = \exp(-2(\frac{1}{2} - b)^2 k) \leqslant \delta$ as soon as $k \geqslant \frac{\ln(1/\delta)}{2(\frac{1}{2}-b)^2}$. We can then obtain a (ε, δ)-estimator for μ from Y for all $\varepsilon \geqslant \max(a - 1, 1 - 1/a)$ and all $\delta > 0$. ◁

■ **Exercise 4 (Pairwise independent random bits).** We will describe a way to generate n pairwise independent uniform random bits $X_1, \ldots, X_n$ using only $\ell = \lceil \log_2 n \rceil$ "true" uniform independent random bits $Y_1, \ldots, Y_\ell$.

▶ ***Question* 4.1)** *Let $(G, \cdot)$ be a finite group, X a random variable over G and U an independent uniform random variable over G. Show that $X \cdot U$ is an uniform random variable over G independent from X.*

Answer. ▷ First, $X \cdot U$ is uniform, indeed for all $g \in G$:

$$\begin{aligned}
\Pr\{\, X \cdot U = g \,\} &= \sum_{g' \in G} \Pr\{\, X = g' \,\} \cdot \Pr\{\, U = {g'}^{-1} \cdot g \,|X = g' \,\} \\
&= \sum_{g' \in G} \Pr\{\, X = g' \,\} \cdot \Pr\{\, U = {g'}^{-1} \cdot g \,\} \\
&= \sum_{g' \in G} \Pr\{\, X = g' \,\} \cdot \frac{1}{|G|} = \frac{1}{|G|},
\end{aligned}$$

since X and U are independent. Furthermore, X and $X \cdot U$ are independent, indeed for all $(g, g') \in G^2$:

$$\begin{aligned}
\Pr\{\, X = g \text{ and } X \cdot U = g' \,\} &= \Pr\{\, X = g \wedge U = g^{-1} \cdot g' \,\} \\
&= \Pr\{\, X = g \,\} \cdot \Pr\{\, U = g^{-1} \cdot g' \,\} \\
&= \Pr\{\, X = g \,\} \cdot \frac{1}{|G|} = \Pr\{\, X = g \,\} \cdot \Pr\{\, X \cdot U = g' \,\},
\end{aligned}$$

since U and $X \cdot U$ are uniform random variables over G. ◁

Let $[i] = \{j : j\text{-th bit of } i \text{ written in binary is } 1\} \subseteq \{1, \dots, \ell\}$ such that $i = \sum_{j \in [i]} 2^{j-1}$ for all $i \in \{1, \dots, n\}$. Consider $Y_1, \dots, Y_\ell$, ℓ uniform independent random bits. We then set $X_i = \bigoplus_{j \in [i]} Y_j$ for $i = 1 \dots n$, where $a \oplus b$ denote the XOR of a and b (i.e. their sum modulo 2). For instance: $13 = 1101$ in binary, thus $X_{13} = Y_4 \oplus Y_3 \oplus Y_1$.

▶ ***Question 4.2)*** *Show that $X_1, \dots, X_n$ are n pairwise independent uniform random bits.*

Answer. ▷ Since for every $i \in \{1, \dots, n\}$, X_i is a non-empty sum of independent uniform random bits, every X_i is an uniform random bit by Question 4.1 with $(G, \cdot) = (\mathbb{Z}_2, +)$.

Let us now prove their pairwise independence. Take $i \neq i' \in \{1, \dots, n\}$; $[i]$ and $[i']$ are two different non-empty sets. Let $A = \bigoplus_{j \in [i] \smallsetminus [i']} Y_j$, $B = \bigoplus_{j \in [i'] \smallsetminus [i]} Y_j$ and $C = \bigoplus_{j \in [i] \cap [i']} Y_j$, such that $X_i = A \oplus C$ and $X_{i'} = B \oplus C$ where A, B, and C are independent random $\{0,1\}$-variables. Since $i \neq i'$, we can assume without loss of generality that $[i] \smallsetminus [i'] \neq \varnothing$. From above, A is thus an uniform random bit. Furthermore since $[i]$ and $[i']$ are two different non-empty sets, $[i'] \smallsetminus [i]$ or $[i] \cap [i']$ is non-empty. Consider first the case where $[i'] \smallsetminus [i] \neq \varnothing$, then B is an uniform random bit, independent from A and C. The same proof as in Question 4.1 shows that $X_i = A \oplus C$ and $X_{i'} = B \oplus C$ are independent uniform random bits, indeed for all $(a, b) \in \{0,1\}^2$:

$$
\begin{aligned}
&\Pr\{ A \oplus C = a \text{ and } B \oplus C = b \} \\
&\qquad = \sum_{c \in \{0,1\}} \Pr\{ C = c\} \cdot \Pr\{A = a \oplus c \text{ and } B = b \oplus c \,|\, C = c \} \\
&\qquad = \sum_{c \in \{0,1\}} \Pr\{ C = c\} \cdot \Pr\{A = a \oplus c \} \cdot \Pr\{ B = b \oplus c \} \\
&\qquad = \sum_{c \in \{0,1\}} \Pr\{ C = c \} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \cdot \frac{1}{2} = \Pr\{ A \oplus C = a \} \cdot \Pr\{ B \oplus C = b \}
\end{aligned}
$$

since A, B and C are independent random variables and $A \oplus C$ and $B \oplus C$ are uniform random bits.

Finally, consider the remaining case where $[i'] \subsetneq [i]$, then A and C are two independent uniform random bits; it follows from Question 4.1 that $X_i = A \oplus C$ and $X_{i'} = C$ are independent. ◁