Unbiased CVM Algorithm

January 9, 2025

1 Notation

For a finite set S, the probability space of uniformly sampling from the set is denoted by $U(S)$, i.e., for each $s \in S$ we have $\mathcal{P}_{U(S)}(s) = |S|^{-1}$. For the bernoulli probability space, over the set $\{0, 1\}$ we'll write $\mathrm{Ber}(p)$, i.e., $P_{\mathrm{Ber}(p)}(\{1\}) = p$. $I(P)$ is the indicator function for a predicate P, i.e., $I(\text{true}) = 1$ and $I(\text{false}) = 0$.

2 Algorithm

Let us fix

- $a_1, \ldots, a_l$ for the stream of elements, and
- n for the maximal number of elements in the buffer, and
- f for the fraction of elements to keep in the buffer, when we subsample.

Note that: nf must be an integer $\frac{1}{2} \leq f < 1$.

Algorithm 1 Unbiased CVM algorithm

```
1:  χ ← ∅; p ← 1
2:  for i ← 1 to l do
3:      χ ← χ − {a_i}
4:      coin ← Ber(p)
5:      if coin then
6:          χ ← χ ∪ {a_i}
7:      end if
8:      if |χ| = n then
9:          χ ← U({S ⊆ χ | |S| = nf})
10:         p ← pf
11:     end if
12: end for
13: return p^{-1}|χ|
```

We will denote the first five lines of the loop (4–7) as step 1, the last four lines (8–11) as step 2, and the distribution of the state of the algorithm after processing i elements of the sequence by Ω_i.

The elements of the probability spaces are pairs composed of a set and a fraction, representing χ and p. For example: $\Omega_0 = U(\{(\emptyset, 1)\})$ is the initial state, $\Omega_1 = U(\{(\{a_1\}, 1)\})$, etc. Ω_l denotes the final state. We introduce χ and p as random variables defined over such probability spaces Ω, in particular, χ (resp. p) is the projection to the first (resp. second) component.

The state of the algorithm after processing only step 1 of the i-th loop iteration is denoted by Ω'_i. So the sequence of states is represented by the distributions $\Omega_0, \Omega'_1, \Omega_1, \ldots, \Omega'_l, \Omega_l$. A few easy-to-see observations without proof for any $0 \leq i \leq l$:

- $p(\omega) \leq 1$ a.s. (almost surely) for $\omega \in \Omega_i$,
- $\chi(\omega) \subseteq \{a_1, \ldots, a_i\}$ a.s. for $\omega \in \Omega_i$,
- $|\chi(\omega)| < n$ a.s. for $\omega \in \Omega_i$.

Also for the intermediate states, we have:

- $p(\omega) \leq 1$ a. s. for $\omega \in \Omega'_i$,
- $\chi(\omega) \subseteq \{a_1, \ldots, a_i\}$ a.s. for $\omega \in \Omega'_i$,
- $|\chi(\omega)| \leq n$ a.s. for $\omega \in \Omega'_i$.

Given a set T with n elements, we'll write $C(T)$ for the nf-subsets of T, i.e.:

$$C(T) = \{\tau \subseteq T \,|\, |\tau| = nf\}.$$

Lemma 1. Let T be a set with n elements, $S \subseteq T$ and $g : \{0, 1\} \to \mathbb{R}_{\geq 0}$ then

$$\int_{U(C(T))} \prod_{s \in S} g(I(s \in \tau)) \leq \prod_{s \in S} \int_{\mathrm{Ber}(f)} g(\tau) d\tau$$

Proof. We'll verify the cases $g(0) \leq g(1)$ and the converse seperately. For the case $g(0) \leq g(1)$: We want to use Theorem 2.11 by Joag-Dev and Proschan [1], which requires us to view $C(T)$ as a permutation distribution.

Let us for that consider $P := U(\{f : T \to \{0, .., |T| - 1\}, f \text{bij.}\})$, i.e., we are assigning randomly distinct positions from 0 to $|T| - 1$ to the elements of T. Then P is clearly a permuation distribution and the random variables describing the positions of each element are negatively associated according to Theorem 2.11.

Now we can represent $C(T)$ as the choice of all the elements whose associated position is greater or equal to $(1 - f)n$, i.e.:

$$\int_{U(C(T))} \prod_{s \in S} g(I(s \in \tau)) = \int_P \prod_{s \in S} g(I(\tau(s) \geq (1 - f)n))\, d\tau := R$$

Now we can use the Property P2 [1, Page 288] since the $\tau(s)$ are negatively associated:

$$R \leq \prod_{s \in S} \int_P g(I(\tau(s) \geq (1 - f)n))\, d\tau \tag{1}$$

Note that this works because $x \to g(I(x \geq (1 - f)n))$ is a monotone increasing function. Moreover the right hand side of Eq. 1 is easy to be seen to be equal to the right hand side of the lemma we want to show.

For the case $g(0) > g(1)$ we can do a similar trick, except in this case we associate with the choice of τ the elements with positions $\{0, \ldots, nf - 1\}$.

Important Note: In the paper by Joag-Dev and Proschan "positive" means non-negative. (See last paragraph of Section 1.) □

Lemma 2. Let $\varphi : (0, 1] \times 0, 1 \to \mathbb{R}_{\geq 0}$ be function, fulfilling the following conditions:

1. $(1 - \alpha)\varphi(x, 0) + \alpha\varphi(x, 1) \leq \varphi(x/\alpha, 1)$ for all $0 < \alpha < 1$, $0 < x \leq 1$, and
2. $\varphi(x, 0) \leq \varphi(y, 0)$ for all $0 < x < y \leq 1$.

Then for all $k \in \{0, \ldots, l\}$, $S \subseteq \{a_1, .., a_k\}$, $\Omega \in \{\Omega_k, \Omega'_k\}$:

$$\mathrm{E}_\Omega \left[\prod_{s \in S} \varphi(p, I(s \in \chi)) \right] \leq \varphi(1, 1)^{|S|}$$

Proof. We show the result using induction over k. Note that we show the statement for arbitrary S, i.e., the induction statements are:

$$P(k) \quad :\leftrightarrow \quad \left(\forall S \subseteq \{a_1, .., a_k\}.\, \mathrm{E}_{\Omega_k} \left[\prod_{s \in S} \varphi(p, I(s \in \chi)) \right] \leq \varphi(1, 1)^{|S|} \right)$$

$$Q(k) \quad :\leftrightarrow \quad \left(\forall S \subseteq \{a_1, .., a_k\}.\, \mathrm{E}_{\Omega'_k} \left[\prod_{s \in S} \varphi(p, I(s \in \chi)) \right] \leq \varphi(1, 1)^{|S|} \right)$$

and we will show $P(0), Q(1), P(1), Q(2), P(2), \ldots, Q(l), P(l)$ successively.

Induction start $P(0)$:
We have $S \subseteq \emptyset$, and hence

$$\mathrm{E}_{\Omega_0} \left[\prod_{s \in S} \varphi(p, I(s \in \chi)) \right] = \mathrm{E}_{\Omega_0} [1] = 1 \leq \varphi(1, 1)^0.$$

Induction step $P(k) \to Q(k + 1)$:
Let $S \subseteq \{a_1, \ldots, a_{k+1}\}$ and define $S' := S - \{a_{k+1}\}$. Note that Ω'_{k+1} can be constructed from Ω_k as a compound distribution, where a_{k+1} is included in the buffer, with the probability p, which is itself a random variable over the space Ω_k.

In particular, for example:

$$\mathcal{P}_{\Omega'_{k+1}}(P(\chi, p)) = \int_{\Omega_k} \int_{\mathrm{Ber}(p(\omega))} P(\chi(\omega) - \{a_{k+1}\} \cup f(\tau), p(\omega))\, d\tau\, d\omega$$

for all predicates P where we define $f(1) = \{a_{k+1}\}$ and $f(0) = \emptyset$.

We distinguish the two cases $a_{k+1} \in S$ and $a_{k+1} \notin S$. If $a_{k+1} \in S$:

$$\begin{aligned}
\mathrm{E}_{\Omega'_{k+1}}\left[\prod_{s\in S}\varphi(p, I(s\in\chi))\right] &= \int_{\Omega_k}\left(\prod_{s\in S'}\varphi(p, I(s\in\chi))\right)\int_{\mathrm{Ber}(p(\omega))}\varphi(p,\tau)\,d\tau\,d\omega \\
&= \int_{\Omega_k}\left(\prod_{s\in S'}\varphi(p, I(s\in\chi))\right)((1-p)\varphi(p,0)+p\varphi(p,1))\,d\omega \\
&\underset{\text{Cond 1}}{\leq} \int_{\Omega_k}\left(\prod_{s\in S'}\varphi(p, I(s\in\chi))\right)\varphi(1,1)\,d\omega \\
&\underset{\text{IH}}{\leq} \varphi(1,1)^{|S'|}\varphi(1,1) = \varphi(1,1)^{|S|}
\end{aligned}$$

If $a_{k+1} \notin S$ then $S' = S$ and:

$$\mathrm{E}_{\Omega'_{k+1}}\left[\prod_{s\in S}\varphi(p, I(s\in\chi))\right] = \int_{\Omega_k}\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \leq_{IH} \varphi(1,1)^{|S'|} = \varphi(1,1)^{|S|} \tag{2}$$

Induction step $Q(k+1) \to P(k+1)$**:**
Let $S \subseteq \{a_1, \ldots, a_{k+1}\}$.

Let us again note that Ω_{k+1} is a compound distribution over Ω_k. In general, for all predicates P:

$$\mathcal{P}_{\Omega_{k+1}}(P(\chi,p)) = \int_{\Omega'_{k+1}} I(|\chi(\omega)|<n)P(\chi(\omega),p(\omega)) + I(|\chi(\omega)|=n)\int_{U(\chi(\omega))}P(\tau, fp(\omega))\,d\tau\,d\omega.$$

With this we can can now verify the induction step:

$$\begin{aligned}
& \mathrm{E}_{\Omega_{k+1}}\left[\textstyle\prod_{s\in S}\varphi(p, I(s\in\chi))\right] \\
= \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|<n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
+ \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|=n)\prod_{s\in S\setminus\chi(\omega)}\varphi(pf,0)\int_{U(C(\chi))}\prod_{s\in S\cap\chi}\varphi(pf, I(s\in\tau))d\tau\,d\omega \\
\underset{\text{Le. 1}}{\leq} \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|<n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
+ \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|=n)\prod_{s\in S\setminus\chi(\omega)}\varphi(pf,0)\prod_{s\in S\cap\chi}\int_{\mathrm{Ber(f)}}\varphi(pf,\tau)d\tau\,d\omega \\
\underset{\text{Cond 2}}{\leq} \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|<n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
+ \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|=n)\prod_{s\in S\setminus\chi(\omega)}\varphi(p,0)\prod_{s\in S\cap\chi}((1-f)\varphi(pf,0)+f\varphi(pf,1))\,d\omega \\
\underset{\text{Cond 1}}{\leq} \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|<n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
+ \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|=n)\prod_{s\in S\setminus\chi(\omega)}\varphi(p,0)\prod_{s\in S\cap\chi}\varphi(p,1)\,d\omega \\
= \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|<n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
+ \;& \textstyle\int_{\Omega'_{k+1}} I(|\chi|=n)\prod_{s\in S}\varphi(p, I(s\in\chi))\,d\omega \\
= \;& \mathrm{E}_{\Omega'_{k+1}}\left[\textstyle\prod_{s\in S}\varphi(p, I(s\in\chi))\right] \underset{\text{IH}}{\leq} \varphi(1,1)^{|S|}
\end{aligned}$$

□

More practical versions of the previous lemma:

Lemma 3. Let $q \leq 1$ and $h : [0, q^{-1}] \to \mathbb{R}_{\geq 0}$ be concave then for all $k \in \{0, \ldots, l\}$, $S \subseteq \{a_1, .., a_k\}$, $\Omega \in \{\Omega_k, \Omega'_k\}$:

$$\mathrm{E}_\Omega \left[\prod_{s \in S} I(p > q) h(p^{-1} I(s \in \chi)) \right] \leq h(1)^{|S|}$$

Proof. Follows from Lemma 2 for $\varphi(p, \tau) := I(p > q)h(\tau p^{-1})$. We just need to check Conditions 1/2, Indeed

$$\begin{aligned} (1-\alpha)\varphi(x,0) + \alpha\varphi(x,1) &= (1-\alpha)I(x>q)h(0) + \alpha I(x>q)h(x^{-1}) \\ &\leq I(x>q)h(\alpha x^{-1}) \leq I(x > q\alpha)h(\alpha x^{-1}) = \varphi(x/\alpha, 1) \end{aligned}$$

for $0 < \alpha < 1$ and $0 < x \leq 1$, where we used that $x > q$ implies $x > q\alpha$. And

$$\varphi(x,0) = I(x>q)h(0) \leq I(y>q)h(0) = \varphi(y,0)$$

for $0 < x < y \leq 1$, where we used that $x > q$ implies $y > q$. □

3 Concentration

This section establishes that the result of the algorithm is concentrated around the cardinality of $A = \{a_1, \ldots, a_l\}$. This will be done by Chernoff bounds for the probability that the estimate is above $(1+\varepsilon)|A|$ (resp. below $(1-\varepsilon)|A|$ assuming p is not too small and a tail estimate for p being too small.

It should be noted that concentration is trivial, if $|A| < n$, i.e., if we never need to do sub-sampling. So we assume $|A| \geq n$.

We define $q := n/(4|A|)$ - notice that $q \leq \frac{1}{4}$.

Let us start with the upper tail bound:

Lemma 4. For any $\Omega \in \{\Omega_0, \ldots, \Omega_l\} \cup \{\Omega'_1, \ldots, \Omega'_l\}$ and $0 < \varepsilon \leq 1$:

$$L := \mathcal{P}_\Omega \left(p^{-1}|\chi| \geq (1+\varepsilon)|A| \wedge p \geq q \right) \leq \exp\left(-\frac{n}{12}\varepsilon^2 \right)$$

Proof. By assumption there exists a k such that $\Omega \in \{\Omega_k, \Omega'_k\}$. Let $A' = A \cap \{a_1, \ldots, a_k\}$.

Moreover, we define

$$\begin{aligned} t &:= q \ln(1+\varepsilon) \\ h(x) &:= 1 + qx(e^{t/q} - 1) \end{aligned}$$

To get a tail estimate, we use the Cramér-Chernoff method:

$$\begin{aligned} L &\underset{t>0}{\leq} \mathcal{P}_\Omega \left(\exp(tp^{-1}|\chi|) \geq \exp(t(1+\varepsilon)|A|) \wedge p \geq q \right) \\ &\leq \mathcal{P}_\Omega \left(I(p \geq q)\exp(tp^{-1}|\chi|) \geq \exp(t(1+\varepsilon)|A|) \right) \\ &\underset{\text{Markov}}{\leq} \exp(-t(1+\varepsilon)|A|)\mathrm{E}_\Omega \left[I(p \geq q)\exp(tp^{-1}|\chi|) \right] \end{aligned}$$

$$
\begin{aligned}
&= \exp(-t(1+\varepsilon)|A|)\mathrm{E}_\Omega\left[\prod_{s\in A'} I(p\geq q)\exp(tp^{-1}I(s\in\chi))\right]\\
&\leq \exp(-t(1+\varepsilon)|A|)\mathrm{E}_\Omega\left[\prod_{s\in A'} I(p\geq q)h(p^{-1}I(s\in\chi))\right]\\
&\underset{\text{Le. 3}}{\leq} \exp(-t(1+\varepsilon)|A|)h(1)^{|A'|}\\
&\underset{h(1)\geq 1}{\leq} (\exp(\ln(h(1))-t(1+\varepsilon)))^{|A|}
\end{aligned}
$$

So we just need to show that (using $|A| = \frac{n}{4q}$):

$$\ln(h(1)) - t(1+\varepsilon) \leq \frac{-q\varepsilon^2}{3}$$

The latter can be established by analyzing the function

$$f(\varepsilon) := -\ln(1+q\varepsilon) + q\ln(1+\varepsilon)(1+\varepsilon) - \frac{q\varepsilon^2}{3} = -\ln(h(1)) + t(1+\varepsilon) - \frac{q\varepsilon^2}{3}.$$

For which it is easy to check $f(0) = 0$ and the derivative with respect to ε is non-negative in the range $0 \leq q \leq 1/4$ and $0 < \varepsilon \leq 1$, i.e., $f(\varepsilon) \geq 0$. □

Using the previous we can estimate bounds for p becoming too small:

Lemma 5.

$$\mathcal{P}_{\Omega_l}(p < q) \leq l\exp\left(-\frac{n}{12}\right)$$

Proof. We'll use a similar strategy as in the Bad_2 bound in the original CVM paper. Let j be maximal, s.t., $q \leq f^j$. Hence $f^{j+1} < q$ and:

$$f^j \leq 2ff^j < 2q = \frac{n}{2|A|}. \tag{3}$$

First, we bound the probability of jumping from $p = f^j$ to $p = f^{j+1}$ at a specific point in the algorithm, e.g., after processing k stream elements. It can only happen if $|\chi| = n$, $p = f^j$ in Ω'_k. Then

$$
\begin{aligned}
\mathcal{P}_{\Omega'_k}(|\chi| \geq n \wedge p = f^j) &\leq \mathcal{P}(p^{-1}|\chi| \geq f^{-j}n \wedge p \geq q)\\
&\underset{\text{Eq. 3}}{\leq} \mathcal{P}(p^{-1}|\chi| \geq 2|A| \wedge p \geq q)\\
&\underset{\text{Le. 4}}{\leq} \exp(-n/12)
\end{aligned}
$$

The probability that this happens ever in the entire process is then at most l times the above which proves the lemma. □

Lemma 6. Let $0 < \varepsilon < 1$ then:

$$L := \mathcal{P}_{\Omega_l}(p^{-1}|\chi| \leq (1-\varepsilon)|A| \wedge p \geq q) \leq \exp\left(-\frac{n}{8}\varepsilon^2\right)$$

Proof. Let us define

$$t := q\ln(1-\varepsilon) < 0$$
$$h(x) := 1 + qx(e^{t/q} - 1)$$

Note that $h(x) \geq 0$ for $0 \leq x \leq q^{-1}$ (can be checked by verifying it is true for $h(0)$ and $h(q^{-1})$ and the fact that the function is affine.)

With these definitions we again follow the Cramér-Chernoff method:

$$\begin{aligned}
L &\underset{t<0}{=} \mathcal{P}_{\Omega_l}\left(\exp(tp^{-1}|\chi|) \geq \exp(t(1-\varepsilon)|A|) \wedge p \geq q\right) \\
&\leq \mathcal{P}_{\Omega_l}\left(I(p \geq q)\exp(tp^{-1}|\chi|) \geq \exp(t(1-\varepsilon)|A|) \wedge p > q\right) \\
&\underset{\text{Markov}}{\leq} \exp(-t(1-\varepsilon)|A|)\mathrm{E}_\Omega\left[I(p \geq q)\exp(tp^{-1}|\chi|)\right] \\
&= \exp(-t(1-\varepsilon)|A|)\mathrm{E}_\Omega\left[\prod_{s\in A} I(p \geq q)\exp(tp^{-1}I(s \in \chi))\right] \\
&\leq \exp(-t(1-\varepsilon)|A|)\mathrm{E}_\Omega\left[\prod_{s\in A} I(p \geq q)h(p^{-1}I(s \in \chi))\right] \\
&\underset{\text{Le. 3}}{\leq} \exp(-t(1-\varepsilon)|A|)(h(1))^{|A|} \\
&= \exp(\ln(h(1)) - t(1-\varepsilon))^{|A|}
\end{aligned}$$

Substituting t and h and using $|A| = \frac{n}{4q}$, we can see that the lemma is true if

$$f(\varepsilon) := -q\ln(1-\varepsilon)(1-\varepsilon) - \ln(1-q\varepsilon) - \frac{q}{2}\varepsilon^2 = t(1-\varepsilon) - \ln(h(1)) - \frac{q}{2}\varepsilon^2$$

is non-negative for $0 \leq q \leq \frac{1}{4}$ and $0 < \varepsilon < 1$. This can be verified by checking that $f(0) = 0$ and that the derivative with respect to ε is non-negative. □

We can now establish the concentration result:

Theorem 1. *Let* $0 < \varepsilon < 1$ *and* $0 < \delta < 1$ *and* $n \geq \frac{12}{\varepsilon^2}\ln\left(\frac{3l}{\delta}\right)$ *then:*

$$L = \mathcal{P}_{\Omega_l}\left(|p^{-1}|\chi| - |A|| \geq \varepsilon|A|\right) \leq \delta$$

Proof. Note that the theorem is trivial if $|A| < n$. If not:

$$\begin{aligned}
L &\leq \mathcal{P}_{\Omega_l}\left(|p^{-1}|\chi| \leq (1-\varepsilon)|A| \wedge p \geq q\right) + \mathcal{P}_{\Omega_l}\left(|p^{-1}|\chi| \geq (1+\varepsilon)|A| \wedge p \geq q\right) + \mathcal{P}_{\Omega_l}\left(p < q\right) \\
&\underset{Le.\ 4-6}{\leq} \exp\left(-\frac{n}{8}\varepsilon^2\right) + \exp\left(-\frac{n}{12}\varepsilon^2\right) + l\exp\left(-\frac{n}{12}\right) \\
&\leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3}
\end{aligned}$$

□

4 Unbiasedness

Let M be large enough such that $p^{-1} \leq M$ a.s. (e.g. we can choose $M = f^{-l}$). Then we can derive from Lemma 3 using $h(x) = x$ and $h(x) = M + 1 - x$ that:

$$\begin{aligned} \mathrm{E}_{\Omega_l}[p^{-1} I(s \in \chi)] &= \mathrm{E}_{\Omega_l}[I(p \geq M^{-1}) p^{-1} I(s \in \chi)] \leq 1 \\ \mathrm{E}_{\Omega_l}[M + 1 - p^{-1} I(s \in \chi)] &= \mathrm{E}_{\Omega_l}[I(p \geq M^{-1})(M + 1 - p^{-1} I(s \in \chi))] \leq M \end{aligned}$$

which implies $\mathrm{E}_{\Omega_l}[p^{-1} I(s \in \chi)] = 1$. By linearity of expectation we can conclude that

$$\mathrm{E}_{\Omega_l}[p^{-1}|\chi|] = \sum_{s \in A} \mathrm{E}_{\Omega_l}[p^{-1} I(s \in \chi)] = |A|.$$

5 Other approaches

Initially I had tried to show that the RV's $I(s \in \chi)$ might be negatively associated. Unfortunately that is wrong, and because of the closure properties of negative association the same is true of $p^{-1} I(s \in \chi)$.

The best counter-example (so far) I have is:

- $a_1 = 1, \ldots, a_{15} = 15$
- $n = 10$
- $f = \frac{1}{2}$

then the RV's $f(\chi) = I(|\chi \cap \{1, .., 10\}| \geq 5)$ and $g(\chi) = |\chi \cap \{11, \ldots, 15\}|$ then

$$Efg > EfEg$$

with

$$\begin{aligned} EfEg &= \frac{59537}{24576} < 2.4225667318 \\ Efg &= \frac{78625}{32256} > 2.4375310019 \end{aligned}$$

Note: I verified this using Haskell with exact computation, i.e, we branch on every coin flip and big integer rational arithmetic. So these are exact values.

References

[1] Kumar Joag-Dev and Frank Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286 – 295, 1983. doi:10.1214/aos/1176346079.