Chapter 5

Sampling, Estimation, and More on the Coupon's Collector Problems II

By Sariel Har-Peled, December 30, 2015[①]

> There is not much talking now. A silence falls upon them all. This is no time to talk of hedges and fields, or the beauties of any country. Sadness and fear and hate, how they well up in the heart and mind, whenever one opens the pages of these messengers of doom. Cry for the broken tribe, for the law and custom that is gone. Aye, and cry aloud for the man who is dead, for the woman and children bereaved. Cry, the beloved country, these things are not yet at an end. The sun pours down on the earth, on the lovely land that man cannot enjoy. He knows only the fear of his heart.
>
> – Alan Paton, Cry, the beloved country.

5.1. Randomized selection – Using sampling to learn the world

5.1.1. Sampling

One of the big advantages of randomized algorithms, is that they sample the world; that is, learn how the input looks like without reading all the input. For example, consider the following problem: We are given a set of U of n objects $u_1, \ldots, u_n$. and we want to compute the number of elements of U that have some property. Assume, that one can check if this property holds, in constant time, for a single object, and let $\psi(u)$ be the function that returns 1 if the property holds for the element u. and zero otherwise. Now, let α be the number of objects in U that have this property. We want to reliably estimate α without computing the property for all the elements of U.

A natural approach, would be to pick a random sample R of m objects, $r_1, \ldots, r_m$ from U (with repetition), and compute $Y = \sum_{i=1}^{m} \psi(r_1)$, and our estimate for α is $\beta = (n/m)Y$. It is natural to ask how far is β from the true estimate.

Lemma 5.1.1. *Let U be a set of n elements, with α of them having a certain property ψ. Let R be a uniform random sample from U (with repetition), and let Y be the number of elements in R that have the property ψ, and let $Z = (n/m)Y$ be the estimate for α. Then, for any $t \geq 1$, we have that*

$$\mathbf{Pr}\left[\alpha - t\frac{n}{2\sqrt{m}} \leq Z \leq \alpha + t\frac{n}{2\sqrt{m}}\right] \geq 1 - \frac{1}{t^2}.$$

Similarly, we have that $\mathbf{Pr}\Big[\mathbf{E}[Y] - t\sqrt{m}/2 \leq Y \leq \mathbf{E}[Y] + t\sqrt{m}/2\Big] \geq 1 - \frac{1}{t^2}$.

Proof: Let $Y_i = \psi(r_i)$ be an indicator variable that is 1 if the ith sample r_i has the property ψ. Now, $Y = \sum_i Y_i$ is a binomial distribution with probability $p = \alpha/n$, and m samples; that is, $Y \sim \mathrm{Bin}(m, p)$. We saw in the previous lecture that, $\mathbf{E}[Y] = mp$, $\mathbf{V}[Y] = mp(1-p)$, and its standard deviation is as such $\sigma_Y = \sqrt{mp(1-p)} \leq \sqrt{m}/2$, as $\sqrt{p(1-p)}$ is maximized for $p = 1/2$. We have $\Delta = \frac{t\sigma_Y n}{m} \leq t\frac{\sqrt{m}n}{2m} = t\frac{n}{2\sqrt{m}}$, since $\sqrt{(\alpha/n)(1-(\alpha/n))}$ is maximized for $\alpha = n/2$. As such,

$$\begin{aligned}\mathbf{Pr}\left[\left|Z - \alpha\right| \geq t\frac{n}{2\sqrt{m}}\right] &\leq \mathbf{Pr}\Big[\left|Z - \alpha\right| \geq \Delta\Big] = \mathbf{Pr}\left[\left|\frac{n}{m}Y - \alpha\right| \geq \Delta\right] = \mathbf{Pr}\left[\left|Y - \frac{m}{n}\alpha\right| \geq \frac{m}{n}\Delta\right] \\ &= \mathbf{Pr}\Big[\left|Y - \mathbf{E}[Y]\right| \geq t\sigma_Y\Big] \leq \frac{1}{t^2},\end{aligned}$$

by Chebychev's inequality. ∎

5.1.1.1. Inverse estimation

We are given a set $U = \{u_1, \ldots, u_n\}$ of n distinct numbers. Let s_i denote the ith smallest number in U – that is s_i is the number of *rank* i in U. We are interested in estimating s_k quickly. So, let us take a sample R of size m. Let $\mathsf{R}_{\leq s_k}$ be the set of all the numbers in R that are $\leq s_k$. For $Y = \left|\mathsf{R}_{\leq s_k}\right|$, we have that $\mu = \mathbf{E}[Y] = km/n$. Furthermore, for any $t \geq 1$, Lemma 5.1.1 implies that $\mathbf{Pr}\Big[\mu - t\sqrt{m}/2 \leq Y \leq \mu + t\sqrt{m}/2\Big] \geq 1 - \frac{1}{t^2}$. In particular, with probability $\geq 1 - 1/t^2$ the number r_- of rank $\ell_- = \Big\lfloor \mu - t\sqrt{m}/2 \Big\rfloor - 1$ in R is smaller than s_k, and similarly, the number r_+ of rank $\ell_+ = \Big\lceil \mu + t\sqrt{m}/2 \Big\rceil + 1$ in R is larger than s_k.

One can conceptually think about the interval $\mathcal{I}(k) = [r_-, r_+]$ as confidence interval – we know that $s_k \in \mathcal{I}(k)$ with probability $\geq 1 - 1/t^2$. But how big is this interval? Namely, how many elements are there in $\mathcal{I}(k) \cap Sample$?

To this end, consider the interval of ranks in the sample that might contain the kth element. By the above, this is

$$\mathcal{I}(k,t) = k\frac{n}{m} + \Big[-t\sqrt{m}/2 - 1, t\sqrt{m}/2 + 1\Big].$$

In particular, consider the maximum $\nu \leq k$, such that $\mathcal{I}(\nu, t)$ and $\mathcal{I}(k,t)$ are disjoint. We have the condition that

$$\nu\frac{n}{m} + t\sqrt{m}/2 + 1 \leq k\frac{n}{m} - t\sqrt{m}/2 - 1 \implies \nu \leq k - t\frac{m^{3/2}}{n} - 1.$$

Setting $g = k - t\frac{m^{3/2}}{n} - 1$ and $h = k + t\frac{m^{3/2}}{n} + 1$, we have that $\mathcal{I}(g,t)$ and $\mathcal{I}(k,t)$ and $\mathcal{I}(h,t)$ are all disjoint with probability $\geq 1 - 3/t^2$.

To this end, let $g = k - \Big\lceil 2\Big(t\frac{n}{2\sqrt{m}}\Big)\Big\rceil$ and $h = k + \Big\lceil 2\Big(t\frac{n}{2\sqrt{m}}\Big)\Big\rceil$. It is easy to verify (using the same argumentation as above) that with probability at least $1 - 3/t^3$, the three confidence $\mathcal{I}(g), \mathcal{I}(k)$ and $\mathcal{I}(h)$ do not intersect. As such, we have

$$\Big|\mathcal{I}(k) \cap \mathsf{R}\Big| \leq h - g \leq 4\left(t\frac{n}{2\sqrt{m}}\right).$$

We thus get the following.

Func **LazySelect**(S , k)
  `Input:` S - set of n elements, k - index of element to be output.
begin
  repeat
    $R \leftarrow \left\{\text{Sample with replacement of } n^{3/4} \text{ elements from } S\right\}$
      $\cup \{-\infty, +\infty\}$.
    Sort R.
    $l \leftarrow \max\left(1, \left\lfloor kn^{-1/4} - \sqrt{n} \right\rfloor\right)$, $h \leftarrow \min\left(n^{3/4}, \left\lfloor kn^{-1/4} + \sqrt{n} \right\rfloor\right)$
    $a \leftarrow R_{(l)}$, $b \leftarrow R_{(h)}$.
    Compute the ranks $r_S(a)$ and $r_S(b)$ of b in S
      `/* using` $2n$ `comparisons */`
    $P \leftarrow \left\{y \in S \;\middle|\; a \le y \le b\right\}$
    `/* done when computing the rank of` a `and` b `*/`
  Until $(r_S(a) \le k \le r_S(b))$ **and** $\left(|P| \le 8n^{3/4} + 2\right)$
  Sort P in $O(n^{3/4} \log n)$ time.
  return $P_{k - r_S(a) + 1}$
end **LazySelect**

Figure 5.1: The **LazySelect** algorithm.

Lemma 5.1.2. *Given a set U of n numbers, a number k, and parameters t and m, one can compute, in $O(m \log m)$ time, two numbers $r_-, r_+ \in U$, such that:*

(A) The number of rank k in U is in the interval $\mathcal{I}[r_-, r_+]$.

(B) There are at most $O\left(tn/\sqrt{m}\right)$ numbers of U in $\mathcal{I}$.

The algorithm succeeds with probability $\ge 1 - 3/t^3$.

Proof: Compute the sample in $O(m)$ time (assuming the input numbers are in an array, say. Next sort the numbers of R in $O(n \log n)$ time, and return the two elements of rank ℓ_- and ℓ_+ in the sorted set, as the boundaries of the interval. The correctness follows from the above discussion. ■

We next use the above observation to get a fast algorithm for selection.

5.1.2. Randomized selection

We are given a set S of n distinct elements, with an associated ordering. For $t \in S$, let $r_S(t)$ denote the rank of t (the smallest element in S has rank 1). Let $S_{(i)}$ denote the ith element in the sorted list of S.

Given k, we would like to compute S_k (i.e., select the kth element). The code of **LazySelect** is depicted in Figure 5.1.

Exercise 5.1.3. Show how to compute the ranks of $r_S(a)$ and $r_S(b)$, such that the expected number of comparisons performed is $1.5n$.

Consider the element $S_{(k)}$ and where it is mapped to in the random sample R. Consider the interval of values

$$\mathrm{I}(j) = \left[R_{(\alpha(j))}, R_{(\beta(j))}\right] = \left\{R_{(k)} \;\middle|\; \alpha(j) \le k \le \beta(j)\right\},$$

where $\alpha(j) = j \cdot n^{-1/4} - \sqrt{n}$ and $\beta(j) = j \cdot n^{-1/4} + \sqrt{n}$.

Lemma 5.1.4. *For a fixed j, we have that* $\mathbf{Pr}\left[S_{(j)} \in \mathrm{I}(j)\right] \geq 1 - 1/(4n^{1/4})$.

Proof: There are two possible bad events: (i) $S_{(j)} < R_{\alpha(j)}$ and (ii) $R_{\beta(j)} < S_{(j)}$. Let X_i be an indicator variable which is 1 if the ith sample is smaller equal to $S_{(j)}$, otherwise 0. We have $p = \mathbf{Pr}[X_i]] = j/n$ and $q = 1 - j/n$. The random variable $X = \sum_{i=1}^{n^{3/4}} X_i$ is the rank of $S_{(j)}$ in the random sample. Clearly, $X \sim B\left(3/4, j/n\right)$ (i.e., X has a binomial distribution with $p = j/n$, and $n^{3/4}$ trials). As such, we have $\mathbf{E}[X] = pn^{3/4}$ and $\mathbf{V}[X] = n^{3/4}pq$.

Now, by Chebyshev inequality

$$\mathbf{Pr}\left[\left|X - pn^{3/4}\right| \geq t\sqrt{n^{3/4}pq}\right] \leq \frac{1}{t^2}.$$

Since $pn^{3/4} = jn^{-1/4}$ and $\sqrt{n^{3/4}(j/n)(1-j/n)} \leq n^{3/8}/2$, we have that the probability of $a > S_{(j)}$ or $b > S_{(j)}$ is

$$\begin{aligned}
\mathbf{Pr}\left[S_{(j)} < R_{\alpha(j)} \text{ or } R_{\beta(j)} < S_{(j)}\right] &= \mathbf{Pr}\left[X < (jn^{-1/4} - \sqrt{n}) \quad \text{or} \quad X > (jn^{-1/4} + \sqrt{n})\right] \\
&= \mathbf{Pr}\left[|X - jn^{-1/4}| \geq 2n^{1/8} \cdot \frac{n^{3/8}}{2}\right] \\
&\leq \frac{1}{(2n^{1/8})^2} = \frac{1}{4n^{1/4}}.
\end{aligned}$$

■

Lemma 5.1.5. **LazySelect** *succeeds with probability* $\geq 1 - O(n^{-1/4})$ *in the first iteration. And it performs only* $2n + o(n)$ *comparisons.*

Proof: By Lemma 5.1.4, we know that $S_{(k)} \in \mathrm{I}(k)$ with probability $\geq 1 - 1/(4n^{1/4})$. This in turn implies that $S_{(k)} \in P$. Thus, the only possible bad event is that the set P is too large. To this end, set $k^- = k - 3n^{3/4}$ and $k^+ = k + 3n^{3/4}$, and observe that, by definition, it holds $\mathrm{I}(k^-) \cap \mathrm{I}(k) = \emptyset$ and $\mathrm{I}(k) \cap \mathrm{I}(k^+) = \emptyset$. As such, we know by Lemma 5.1.4, that $S_{(k^-)} \in \mathrm{I}(k^-)$ and $S_{(k^+)} \in \mathrm{I}(k^+)$, and this holds with probability $\geq 1 - \frac{2}{4n^{1/4}}$. As such, the set P, which is by definition contained in the range $\mathrm{I}(k)$, has only elements that are larger than $S_{(k^-)}$ and smaller than $S_{(k^+)}$. As such, the size of P is bounded by $k^+ - k^- = 6n^{3/4}$. Thus, the algorithm succeeds in the first iteration, with probability $\geq 1 - \frac{3}{4n^{1/4}}$.

As for the number of comparisons, an iteration requires

$$O(n^{3/4}\log n) + 2n + O(n^{3/4}\log n) = 2n + o(n)$$

comparisons ■

Any deterministic selection algorithm requires $2n$ comparisons, and **LazySelect** can be changed to require only $1.5n + o(n)$ comparisons (expected).

5.2. The Coupon Collector's Problem Revisited

5.2.1. Some technical lemmas

Unfortunately, in Randomized Algorithms, many of the calculations are awful[②]. As such, one has to be dexterous in approximating such calculations. We present quickly a few of these estimates.

[②] "In space travel," repeated Slartibartfast, "all the numbers are awful." – Life, the Universe, and Everything Else, Douglas Adams.

Lemma 5.2.1. *For $x \geq 0$, we have $1 - x \leq \exp(-x)$ and $1 + x \leq e^x$. Namely, for all x, we have $1 + x \leq e^x$.*

Proof: For $x = 0$ we have equality. Next, computing the derivative on both sides, we have that we need to prove that $-1 \leq -\exp(-x) \iff 1 \geq \exp(-x) \iff e^x \geq 1$, which clearly holds for $x \geq 0$.

A similar argument works for the second inequality. ■

Lemma 5.2.2. *For any $y \geq 1$, and $|x| \leq 1$, we have $\left(1 - x^2\right)^y \geq 1 - yx^2$.*

Proof: Observe that the inequality holds with equality for $x = 0$. So compute the derivative of x of both sides of the inequality. We need to prove that

$$y(-2x)\left(1 - x^2\right)^{y-1} \geq -2yx \iff \left(1 - x^2\right)^{y-1} \leq 1,$$

which holds since $1 - x^2 \leq 1$, and $y - 1 \geq 0$. ■

Lemma 5.2.3. *For any $y \geq 1$, and $|x| \leq 1$, we have $\left(1 - x^2 y\right)e^{xy} \leq (1 + x)^y \leq e^{xy}$.*

Proof: The right side of the inequality is standard by now. As for the left side. Observe that

$$(1 - x^2)e^x \leq 1 + x,$$

since dividing both sides by $(1 + x)e^x$, we get $1 - x \leq e^{-x}$, which we know holds for any x. By Lemma 5.2.2, we have

$$\left(1 - x^2 y\right)e^{xy} \leq \left(1 - x^2\right)^y e^{xy} = \left(\left(1 - x^2\right)e^x\right)^y \leq \left(1 + x\right)^y \leq e^{xy}.$$ ■

5.2.2. Back to the coupon collector's problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

In the previous lecture, we showed that

$$\mathbf{Pr}\left[\#\text{ of trials} \geq n \log n + n + t \cdot n\frac{\pi}{\sqrt{6}}\right] \leq \frac{1}{t^2},$$

for any t.

A stronger bound, follows from the following observation. Let Z_i^r denote the event that the ith coupon was not picked in the first r trials. Clearly,

$$\mathbf{Pr}\left[Z_i^r\right] = \left(1 - \frac{1}{n}\right)^r \leq \exp\left(-\frac{r}{n}\right).$$

Thus, for $r = \beta n \log n$, we have $\mathbf{Pr}\left[Z_i^r\right] \leq \exp\left(-\frac{\beta n \log n}{n}\right) = n^{-\beta}$. Thus,

$$\mathbf{Pr}\left[X > \beta n \log n\right] \leq \mathbf{Pr}\left[\bigcup_i Z_i^{\beta n \log n}\right] \leq n \cdot \mathbf{Pr}\left[Z_1\right] \leq n^{-\beta+1}.$$

Lemma 5.2.4. *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then, we have* $\mathbf{Pr}\Big[X > n\ln n + cn\Big] \leq e^{-c}$.

Proof: The probability we fail to pick the first type of coupon is $\alpha = (1-1/n)^m \leq \exp\Big(-\frac{n\ln n+cn}{n}\Big) = \exp(-c)/n$. As such, using the union bound, the probability we fail to pick all n types of coupons is bounded by $n\alpha = \exp(-c)$, as claimed. ■

In the following, we show a slightly stronger bound on the probability, which is $1-\exp(-e^{-c})$. To see that it is indeed stronger, observe that $e^{-c} \geq 1-\exp(-e^{-c})$.

5.2.3. An asymptotically tight bound

Lemma 5.2.5. *Let $c > 0$ be a constant, $m = n\ln n + cn$ for a positive integer n. Then for any constant k, we have* $\lim_{n\to\infty}\binom{n}{k}\Big(1-\frac{k}{n}\Big)^m = \frac{\exp(-ck)}{k!}$.

Proof: By Lemma 5.2.3, we have

$$\left(1-\frac{k^2m}{n^2}\right)\exp\left(-\frac{km}{n}\right) \leq \left(1-\frac{k}{n}\right)^m \leq \exp\left(-\frac{km}{n}\right).$$

Observe also that $\lim_{n\to\infty}\left(1-\frac{k^2m}{n^2}\right) = 1$, and $\exp\left(-\frac{km}{n}\right) = n^{-k}\exp(-ck)$. Also,

$$\lim_{n\to\infty}\binom{n}{k}\frac{k!}{n^k} = \lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{n^k} = 1.$$

Thus, $\lim_{n\to\infty}\binom{n}{k}\Big(1-\frac{k}{n}\Big)^m = \lim_{n\to\infty}\frac{n^k}{k!}\exp\left(-\frac{km}{n}\right) = \lim_{n\to\infty}\frac{n^k}{k!}n^{-k}\exp(-ck) = \frac{\exp(-ck)}{k!}$. ■

Theorem 5.2.6. *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then, for any constant $c \in \mathbb{R}$, and $m = n\ln n + cn$, we have* $\lim_{n\to\infty}\mathbf{Pr}\Big[X > m\Big] = 1-\exp\Big(-e^{-c}\Big)$.

Before dwelling into the proof, observe that $1-\exp(-e^{-c}) \approx 1-(1-e^{-c}) = e^{-c}$. Namely, in the limit, the upper bound of Lemma 5.2.4 is tight.

Proof: We have $\mathbf{Pr}\Big[X > m\Big] = \mathbf{Pr}\Big[\cup_i Z_i^m\Big]$. By inclusion-exclusion, we have

$$\mathbf{Pr}\left[\bigcup_i Z_i^m\right] = \sum_{i=1}^{n}(-1)^{i+1}P_i^n,$$

where $P_j^n = \sum_{1\leq i_1<i_2<\ldots<i_j\leq n}\mathbf{Pr}\left[\bigcap_{v=1}^{j}Z_{i_v}^m\right]$. Let $S_k^n = \sum_{i=1}^{k}(-1)^{i+1}P_i^n$. We know that $S_{2k}^n \leq \mathbf{Pr}\Big[\bigcup_i Z_i^m\Big] \leq S_{2k+1}^n$.

By symmetry,

$$P_k^n = \binom{n}{k}\mathbf{Pr}\left[\bigcap_{v=1}^{k}Z_v^m\right] = \binom{n}{k}\Big(1-\frac{k}{n}\Big)^m,$$

Thus, $P_k = \lim_{n\to\infty} P_k^n = \exp(-ck)/k!$, by Lemma 5.2.5. Thus, we have

$$S_k = \sum_{j=1}^{k} (-1)^{j+1} P_j = \sum_{j=1}^{k} (-1)^{j+1} \cdot \frac{\exp(-cj)}{j!}.$$

Observe that $\lim_{k\to\infty} S_k = 1 - \exp(-e^{-c})$ by the Taylor expansion of $\exp(x)$ (for $x = -e^{-c}$). Indeed,

$$\exp\Big(x\Big) = \sum_{j=0}^{\infty} \frac{x^j}{j!} = \sum_{j=0}^{\infty} \frac{(-e^{-c})^j}{j!} = 1 + \sum_{j=1}^{\infty} \frac{(-1)^j \exp(-cj)}{j!}.$$

Clearly, $\lim_{n\to\infty} S_k^n = S_k$ and $\lim_{k\to\infty} S_k = 1 - \exp(-e^{-c})$. Thus, (using fluffy math), we have

$$\lim_{n\to\infty} \mathbf{Pr}\Big[X > m\Big] = \lim_{n\to\infty} \mathbf{Pr}\Big[\cup_{i=1}^{n} Z_i^m\Big] = \lim_{n\to\infty} \lim_{k\to\infty} S_k^n = \lim_{k\to\infty} S_k = 1 - \exp(-e^{-c}).$$ ■