Asymptotics for a geometric coupon collector process on $\mathbb{N}$

Jacob Richey

February 11, 2021

1 Introduction & Results

Consider the following coupon collector process on $\mathbb{N} = \{0, 1, \ldots\}$. Let $(X_i)_{i\in\mathbb{N}}$ be iid with $\mathrm{Geo}(\alpha)$ distribution, where $\alpha \in (0, 1)$ is the failure probability, i.e. for $v \in \mathbb{N}$,

$$\mathbb{P}(X = v) = p_v = (1 - \alpha)\alpha^v. \tag{1.1}$$

Let V_n be the set of values collected by the X_i up to X_n, i.e.

$$V_n = \{X_1, X_2, \ldots, X_n\}, \tag{1.2}$$

viewed as a set. For example, if $X_1 = 1, X_2 = 4, X_3 = 1, X_4 = 5$, then $V_4 = \{1, 4, 5\}$. Also define the following statistic over V_n, for which we seek a limit theorem:

$$L_n = \sum_{v\in V_n} v^2. \tag{1.3}$$

This sum can be analyzed to give asymptotic formulas the mean and variance of L_n, which are accurate up to a negligible error in n. One obtains $\mathbb{E}L_n \sim \log^3 n$ and $\operatorname{Var} L_n \sim \log^4 n$. (The exact constants appear to be difficult to compute.) Unfortunately, these computations, along with Feller's theorem, also show that there is no central limit theorem for L_n, i.e.

$$\frac{L_n - \mathbb{E}L_n}{\sqrt{\operatorname{Var} L_n}} \tag{1.4}$$

does *not* converge to a normal random variable. (This can be proved rigorously, using a condition called 'uniform asymptotic negligibility.' Interestingly, if the v^2 in the definition of L_n is replaced by $v^{2-\delta}$ for any $\delta > 0$, a CLT would hold.) Let $m_n = \mathbb{E}[\max V_n] \approx c_\alpha \log n$. Then the asymptotic distribution of L_n is roughly

$$L_n \approx m_n^3/3 + D \cdot m_n^2 + o(m_n^2), \tag{1.5}$$

where

$$D = \#\{v \in V_n : v > m_n\} - \#\{v \in V_n : v < m_n\} \tag{1.6}$$

is approximately a difference of two independent Poisson random variables. Roughly speaking, L_n can be thought of as a sum of iid variables. To see this, we will couple with the following iid process. For $n = 1, 2, \ldots$ and $v \in \mathbb{N}$, define **independent** Bernoulli ($\{0, 1\}$-valued) variables $A_v^{(n)}$ with

$$\mathbb{P}(A_v^{(n)} = 1) = 1 - (1 - p_v)^n = \mathbb{P}(v \in V_n) = \mathbb{E}Z_v^{(n)}. \tag{1.7}$$

The A's are associated to their own 'coupon collector process' $\widetilde{V_n}$:

$$\widetilde{V_n} = \{v \in \mathbb{N} : A_v^{(n)} = 1\}. \tag{1.8}$$

We are interested in the statistic

$$\widetilde{L_n} = \sum_{v \in \widetilde{V_n}} v^2 = \sum_{v \in \mathbb{N}} v^2 A_v^{(n)}. \tag{1.9}$$

We prove the following, where TV denotes total variation distance between probability distributions:

Theorem 1.1. *$\mathrm{TV}(L_n, \widetilde{L_n}) \to 0$ as $n \to \infty$.*

This implies that L_n is essentially a sum of iid random variables $A_v^{(n)}$ taking two possible values, so L_n is a relatively simple object.

2 Outline & Proofs

Define the maximum variables

$$M_n = \max V_n, \qquad \widetilde{M_n} = \max \widetilde{V_n}. \tag{2.1}$$

We present two lemmas that describe the behavior of M_n. Roughly speaking, M_n is sharply concentrated around $\sim c_\alpha \log n$, where $c_\alpha = \frac{1}{-\log \alpha}$.

Lemma 2.1. *As $n \to \infty$,*

$$\mathbb{P}([c_\alpha \log n - \sqrt{\log n}] \not\subset V_n) \to 0. \tag{2.2}$$

and similarly for $\widetilde{V_n}$.

Proof. For V_n (the argument for $\widetilde{V_n}$ is identical), since $(1-p_v)^n$ is a strictly increasing function of v, a union bound gives

$$\begin{aligned}
\mathbb{P}([c_\alpha \log n - \sqrt{\log n}] \not\subset V_n) &\leq \sum_{v=0}^{c_\alpha \log n - \sqrt{\log n}} \mathbb{P}(v \notin V_n) & (2.3)\\
&\leq (c_\alpha \log n - \sqrt{\log n} + 1)(1 - p_{c_\alpha \log n - \sqrt{\log n}})^n & (2.4)\\
&= (c_\alpha \log n - \sqrt{\log n} + 1)(1 - (1-\alpha)\frac{\alpha^{-\sqrt{\log n}}}{n})^n & (2.5)\\
&\leq (c_\alpha \log n - \sqrt{\log n} + 1)\exp(-(1-\alpha)\alpha^{-\sqrt{\log n}}) & (2.6)
\end{aligned}$$

□

Lemma 2.2. *As $n \to \infty$,*

$$\mathbb{P}(M_n \geq c_\alpha \log n + \sqrt{\log n}) \to 0 \tag{2.7}$$

and similarly for $\widetilde{M_n}$.

Proof. We compute explicitly with the distribution of M_n – the idea for $\widetilde{M_n}$ is similar. Note that

$$\mathbb{P}(M_n \geq k) = 1 - \mathbb{P}(X < k)^n \tag{2.8}$$
$$= 1 - \left(1 - \alpha^k\right)^n . \tag{2.9}$$

Plugging in $k = c_\alpha \log n + \sqrt{\log n}$ and approximating asymptotically with exponentials gives

$$\mathbb{P}(M_n \geq c_\alpha \log n + \sqrt{\log n}) \leq 2\alpha^{\sqrt{\log n}}. \tag{2.10}$$

□

2.1 Coupling

The remainder of this note is devoted to showing that $\widetilde{V_n}$ and V_n are close in distribution, which implies that L_n and $\widetilde{L_n}$ are also close in distribution, since one applies the same function to get from V_n to L_n as to get from $\widetilde{V_n}$ to $\widetilde{L_n}$. To do so, we explicitly couple V_n and $\widetilde{V_n}$ on the same probability space, and show that the two models agree with high probability. To do so, we construct V_n by 'checking values backwards from ∞.' Fix n, and for $v > 1$, let

$$S_v = \{t \leq n : X_t = v\} \tag{2.11}$$

be the set of indices in $[n]$ taking value v and let

$$\widetilde{S}_v = p_v \text{ percolation on } [n], \tag{2.12}$$

i.e. $t \in \widetilde{S}_v$ with probability p_v for each t and v all independently. The key idea is to construct S_v one value at a time, starting with the largest values. Given $\{S_w : w > v\}$, the distribution of S_v is p'_v percolation on $[n] \setminus \bigcup_{w>v} S_w$, where

$$p'_v = \frac{p_v}{p_0 + p_1 + \cdots + p_v}. \tag{2.13}$$

(Since $n < \infty$, this is a well-defined construction: there will be a 'random' starting value, namely $v = M_n$, with $S_w = 0$ for $w > M_n$.)

Proposition 2.3. *There exists a coupling between the sequences $(S_v)_v$ and $(S'_v)_v$ such that $S_v = S'_v$ with high probability as $n \to \infty$.*

This implies the main theorem, since V_n and $\widetilde{V_n}$ are obtained in the same way from $(S_v)_v$ and $(S'_v)_v$, respectively.

Proof. Note that for ε sufficiently small depending on α, for $v \geq (1-\varepsilon)m_n$,

$$|p_v - p'_v| \leq C_\varepsilon p_v^2 < n^{-1-\delta} \tag{2.14}$$

for some $\delta > 0$. Thus, we can couple S_v and S'_v (for $v > (1-\varepsilon)m_n$) so that we have the following (crude) bound:

$$\mathbb{E}\left[S_v \Delta S'_v | \{S_w : w > v\}\right] \leq p_v \left| \bigcup_{w>v} S_w \right| + (p_v - p'_v)n. \tag{2.15}$$

(The first term comes from the fact that S'_v is sampled independently, so there are $\bigcup_{w>v} S_j$ additional chances to roll value v for S'_v that have already been used up for S_v. The second term corresponds to the remaining indices, of which there are at most n.) Taking expectations, using the bounds 2.14 and $\mathbb{E}\left|\bigcup_{w>v} S_w\right| \leq p_v n$ gives

$$\mathbb{E}[S_v \Delta S'_v] \leq C p_v^2 n \leq n^{-\delta} \tag{2.16}$$

for some $\delta > 0$. Finally, summing over at most $C \log n$ values v by Lemma 2.2, we get that

$$\mathbb{P}(S_v = S'_v \text{ for } v > (1-\varepsilon)m_n)) = 1 - o(1). \tag{2.17}$$

By Lemma 2.1, the same holds with high probability for all values $v \leq (1-\varepsilon)m_n$, so that $S_v = S'_v$ for all v with high probability, completing the proof. □

Question 2.4. *The same type of argument should hold under very mild assumptions on the distribution p_v. What do we need exactly?*