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

Jacob Richey

March 29, 2023

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 sufficiently light tailed distribution. This note focuses on any stretched exponential distribution, i.e. for some $\alpha > 0$, $v \in \mathbb{N}$,

$$\mathbb{P}(X = v) = p_v = C_\alpha \exp(-\alpha\sqrt{v}), \tag{1.1}$$

though similar results should hold for a large class of sufficiently light-tailed distributions. 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 $V_4 = \{1, 4, 5\}$. We seek a limit description of statistics like

$$L_n = \sum_{v\in V_n} \sqrt{v} = \sum_{v\in\mathbb{N}} \sqrt{v} 1\{X_i = v \text{ for some } i \in [n]\} := \sum_{v\in\mathbb{N}} \sqrt{v} A_v^n. \tag{1.3}$$

Note that L is not a sum of independent random variables, since the indicators A_v^n are not independent (though they are only 'midly' dependent for large n.) Our main aim in this note is to show that statistics like L_n are close in distribution to an iid sum in such a way that we can do computations. The strategy is is to couple with the following iid process. For $n = 1, 2, \ldots$ and $v \in \mathbb{N}$, let $\widetilde{A}_v^n$ be independent Bernoullis with

$$\mathbb{P}(\widetilde{A}_v^n = 1) = \mathbb{E}[A_v^n] = 1 - (1 - p_v)^n. \tag{1.4}$$

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

$$\widetilde{V_n} = \{v \in \mathbb{N} : \widetilde{A}_v^n = 1\}, \tag{1.5}$$

and corresponding statistic

$$\widetilde{L_n} = \sum_{v\in\widetilde{V_n}} \sqrt{v} = \sum_{v\in\mathbb{N}} \sqrt{v}\widetilde{A}_v^n. \tag{1.6}$$

Our main result says that V and $\widetilde{V}$ are asymptotically identical in distribution. Here d_{TV} is the total variation between random variables Y, Z, given by

$$d_{TV}(Y, Z) = \inf_\pi \pi\{(y, z) : y \neq z\}, \tag{1.7}$$

where the infemum is taken over all couplings π of X and Y, i.e. all probability measures $\pi = (\pi_Y, \pi_Z)$ on $\Omega \times \Omega$ with marginal distributions Y and Z.

Theorem 1.1. *For any $\delta > 0$, $n^{1-\delta} d_{TV}(V_n, \widetilde{V_n}) \to 0$ as $n \to \infty$.*

This says that the distribution of V is well appoximated by that of $\widetilde{V}$. Thus we obtain that any statistics like L and $\widetilde{L}$ built from V or $\widetilde{V}$ in the same way satisfy the same total variation bound. We do not get a bound on something like $\mathbb{E}[L - \widetilde{L}]$ for free, because on the set where L and $\widetilde{L}$ disagree under the optimal coupling π they could be very large. Conveniently, we work with a coupling π that has V and $\widetilde{V}$ independent conditionally on containing unusually large values, so we can show:

Corollary 1.2. *Let π be the coupling defined in Section 3. For any $\gamma \geq 1$, any $\delta > 0$ and all sufficiently large n,*

$$\mathbb{E}_\pi \left[|L_n - \widetilde{L}_n|^\gamma \right] \leq n^{-1+\delta}. \tag{1.8}$$

In particular, $\left| \mathbb{E}[L_n^\gamma] - \mathbb{E}[\widetilde{L}_n^\gamma] \right| = o_n(1).$

We can use this corollary to do near exact computations for L using $\widetilde{L}$. The expectations are the same for both: we have

$$\mathbb{E} L_n = \mathbb{E}\widetilde{L}_n = \sum_v \sqrt{v}(1 - (1 - p_v)^n) = \frac{2}{3\alpha^3} (\log n)^3 + o((\log n)^3). \tag{1.9}$$

The variance is order $(\log n)^4$:

$$\operatorname{Var} \widetilde{L}_n = \sum_v \operatorname{Var}(\sqrt{v}\widetilde{A}_v) \sim \sum_{v \geq \epsilon (\log n)^2} v \cdot \operatorname{Var}(\widetilde{A}_v) = \Theta((\log n)^4), \tag{1.10}$$

and by the corollary $\operatorname{Var} L_n = \operatorname{Var} \widetilde{L}_n + o(1)$. But Theorem 1.1 allows us to get much more precise distributional information. Let $v_n = \frac{1}{\alpha^2}(\log n)^2$, and decompose L_n as

$$L_n = \sum_{v \leq v_n} \sqrt{v} + \sum_{v \geq v_n} \sqrt{v} A_v^n - \sum_{v < v_n} \sqrt{v} (A_v^n)^c = \frac{3}{2} v_n^{3/2} + L_n^+ - L_n^-. \tag{1.11}$$

Applying the Lindeberg-Feller CLT to $\widetilde{L}^+$ and $\widetilde{L}^-$, using Theorem 1.1 along with Lemmas 2.1 and 2.2 gives the following description.

Corollary 1.3. *There exist constants $\mu^\pm$ and $\sigma^\pm$ such that we have the distributional convergences*

$$\frac{L_n^\pm - \mu^\pm (\log n)^2}{\sigma^\pm \log n} \to_d \mathcal{N}(0, 1). \tag{1.12}$$

In other words, we have the approximate distributional equality

$$\boxed{L_n \approx_d \frac{3}{2\alpha^3} (\log n)^3 + (\mu^+ - \mu^-)(\log n)^2 + Z \log n} \tag{1.13}$$

where Z is normal with mean 0 and variance $(\sigma^+)^2 + (\sigma^-)^2$. In particular, we have the almost sure convergences

$$\frac{L_n}{(\log n)^3} \to_{a.s.} \frac{3}{2\alpha^2} \tag{1.14}$$

and

$$\frac{L_n - \frac{3}{2\alpha^2}(\log n)^3}{(\log n)^2} \to_{a.s.} \mu^+ - \mu^- \tag{1.15}$$

(I write the approximate equality 1.13 this way for brevity – a precise statement would be that the total variation between the LHS and RHS converges to 0 as $n \to \infty$.) The constants are somewhat explicit, depending only on α, in terms of some integrals:

$$\mu^+ = \lim_{n\to\infty} \frac{\mathbb{E}L_n^+}{(\log n)^2} = \frac{1}{\alpha}\int_0^\infty q(z)\,dz \tag{1.16}$$

$$\mu^- = \lim_{n\to\infty} \frac{\mathbb{E}L_n^-}{(\log n)^2} = \frac{1}{\alpha}\int_0^\infty 1 - q(-z)\,dz \tag{1.17}$$

$$(\sigma^+)^2 = \lim_{n\to\infty} \frac{\mathrm{Var}\, L_n^+}{(\log n)^2} = \frac{1}{\alpha}\int_0^\infty q(z)(1-q(z))\,dz \tag{1.18}$$

$$(\sigma^-)^2 = \lim_{n\to\infty} \frac{\mathrm{Var}\, L_n^-}{(\log n)^2} = \frac{1}{\alpha}\int_0^\infty q(-z)(1-q(-z))\,dz \tag{1.19}$$

where for $z \in \mathbb{R}$,

$$q(z) = \lim_{n\to\infty} \mathbb{E}A^n_{v_n + z\log n} = \exp\left(-C_\alpha \exp\left(\frac{1}{2}\alpha^2 z\right)\right) \tag{1.20}$$

(I omit the proof, which is easy – just compute the expectations and variances of L^+ and L^-, then apply the CLT – but involves a lot of annoying error terms, since it requires truncating the sums at ℓ_n and r_n. The bulk contribution to those expectations and variances come from values $v = v_n + z\log n$ for fixed z. If $z = \pm\omega_n(1)$ then the contribution of v to $L^\pm$ is lower order.)

Aside: It appears that $\mu^+ = \mu^-$ for exactly one value of α, namely $\alpha \approx 1.371$. Is there any significance of this value of α?

2 Preliminaries

Define the maximum variables

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

We start with two lemmas that describe the tail of the p_v distribution. Recall that $p_v \sim \exp(-\alpha\sqrt{v})$. In a nutshell, V_n contains all values up to just under $\frac{1}{\alpha^2}(\log n)^2$, and no values just above that point. Note that at that value we have $p_{\alpha^{-2}(\log n)^2} \sim n^{-1}$, i.e. the expected number of occurrences of values $v \approx \alpha^{-2}(\log n)^2$ among the X_i's is $\Theta(1)$.

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

$$\mathbb{P}(\{1, 2, \ldots, \lfloor \alpha^{-2}(\log n)^2 - (\log n)^{3/2}\rfloor\} \not\subset V_n) \to 0. \tag{2.2}$$

and similarly for $\widetilde{V_n}$.

Proof. Write $\ell_n = \lfloor \alpha^{-2}(\log n)^2 - (\log n)^{3/2} \rfloor$, and note the Taylor approximation

$$\sqrt{\ell_n} \approx \alpha^{-1}\log n - \frac{1}{2}\alpha(\log n)^{1/2}, \tag{2.3}$$

where the approximation symbol means we have upper and lower bounds by constants. By a union bound and some algebra (and ignoring irrelevant constants),

$$\mathbb{P}([\ell_n] \not\subset V_n) \leq \sum_{v \leq \ell_n} \mathbb{P}(v \notin V_n) \leq (\log n)^2(1-p_{\ell_n})^n \leq (\log n)^2 \exp(-C_\alpha \exp(\alpha^2(\log n)^{1/2})) \to 0. \tag{2.4}$$

□

We chose $(\log n)^{3/2}$ here so that 1) the contribution of the segment $[\ell_n, \alpha^{-2}(\log n)^2]$ is smaller order than the bulk – for $g(v) = \sqrt{v}$, the bulk is order $(\log n)^3$, while values in that interval contribute at most $\sqrt{(\log n)^2} \cdot (\log n)^{3/2} = (\log n)^{5/2}$ – and 2) the above probability converges to 0.

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

$$\mathbb{P}(M_n \geq \alpha^{-2}(\log n)^2 + (\log n)^{3/2}) \to 0 \tag{2.5}$$

and similarly for $\widetilde{M_n}$.

Proof. Similar to Lemma 2.1. Let $r_n = \lfloor \alpha^{-2}(\log n)^2 + (\log n)^{3/2} \rfloor$. Here we need to sum the tail of our stretched exponential, which is do-able by comparing with an integral:

$$\mathbb{P}(X > v) = \sum_{w > v} p_w \approx \sqrt{v}\exp(-\alpha\sqrt{v}). \tag{2.6}$$

By a Taylor approximation for r_n, and more algebra with exponentials,

$$\mathbb{P}(M_n > r_n) = 1 - \mathbb{P}(X \leq r_n)^n = O(\log n \exp(-\sqrt{\log n})) \to 0. \tag{2.7}$$

□

We will also need the following basic fact about binomial distributions:

Fact 2.3. *Let $B \sim$ Binomial(n,p), $B' \sim$ Binomial(n,q), and $B'' \sim$ Binomial(m,q). Then*

$$d_{TV}(B, B'') \leq d_{TV}(B, B') + d_{TV}(B', B'') \leq n|p-q| + |n-m|q \tag{2.8}$$

In particular, there exists a coupling between B and B'' such that $\mathbb{P}(B \neq B') \leq n|p-q| + |n-m|q$.

These crude bounds from coupling B, B', and B'' in the obvious way (i.e. using the same Bernoullis for all three), then using Markov's inequality to bound $\mathbb{P}(B \neq B')$ or $\mathbb{P}(B \neq B'')$. (These may even be the correct orders for the TV if $|p-q|$ and $|n-m|$ are small, I can't find a reference but surely it's written up somewhere.)

3 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. The construction works by 'adding values backwards from ∞.' Fix n, and for $v > 1$, let

$$S_v = \{t \leq n : X_t = v\} \tag{3.1}$$

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

$$\widetilde{S}_v = p_v \text{ percolation on } [n], \tag{3.2}$$

i.e. $t \in \widetilde{S}_v$ with probability p_v for each t and v all independently. Note that V_n is a measurable function of $(S_v)_v$, namely

$$V_n = \{v : S_v \neq \emptyset\}, \tag{3.3}$$

and similarly for $\widetilde{V}_n$. We now define the coupling between the sequences (S_v) and $(\widetilde{S}_v)$, i.e. a construction of the pair $((S_v)_v, (\widetilde{S}_v)_v)$ on a single probability space, so that the marginals agree with the definitions just given. The coupling is constructed recursively as follows:

- Start with $S_v = \widetilde{S}_v = \emptyset$ for $v > r_n$ (recall $r_n = \lfloor \alpha^{-2}(\log n)^2 + (\log n)^{3/2} \rfloor$ as in the proof of Lemma 2.2) with probability $\mathbb{P}(M_n < r_n)$. With the complementary probability, generate the full sequence (S_v) conditionally on $M_n \geq r_n$ and generate $(\widetilde{S}_v)$ independently. (The latter case won't matter because it has small probability.)

- Given all the sets S_w for $w > v$, generate S_v by adding each $i \in n \setminus \bigcup_{w>v} S_w$ to S_v independently with probability

$$p'_v = \frac{p_v}{p_0 + p_1 + \cdots + p_v} \tag{3.4}$$

  Note that conditionally on $(S_w)_{w>v}$, $|S_v|$ has Binomial$(n - \left|\bigcup_{w>v} S_w\right|, p'_v)$ distribution.

- Use the coupling guaranteed by 2.3 to generate $|\widetilde{S}_v|$ using $|S_v|$, so that $|\widetilde{S}_v|$ has Binomial(n, p_v) distribution. Then if $|\widetilde{S}_v| = |S_v|$, set $\widetilde{S}_v = S_v$, and otherwise choose the indices for $\widetilde{S}_v$ independently.

Note that this coupling has the correct marginals, i.e. the S_v and $\widetilde{S}_v$ constructed this way give rise to the same distribution for V_n and $\widetilde{V_n}$ described at the beginning of this note. We now turn to the central proposition:

Proposition 3.1. *The above coupling has the property that $S_v = \widetilde{S}_v$ for all $v \geq \ell_n$ with high probability as $n \to \infty$.*

This will enable us to do computations for L_n using $\widetilde{L}_n$ instead, since V_n and $\widetilde{V_n}$ are obtained in the same way from $(S_v)_v$ and $(\widetilde{S}_v)_v$, respectively.

Proof. Let $\mathbb{P}_v = \mathbb{P}[\cdot|(S_w)_{w>v}]$ denote the conditional expectation given the history of the coupling. The definition of the coupling and Fact 2.3 give

$$\mathbb{P}_v[S_v \neq \widetilde{S}_v] \leq \mathbb{P}(M_n > r_n) + (p'_v - p_v)n + \left|\bigcup_{w>v} S_w\right| p_v. \tag{3.5}$$

Some algebra shows $p'_v - p_v \leq C\sqrt{\ell_n}p^2_{\ell_n}$ for $v \geq \ell_n$. Also, observe that

$$\mathbb{E}\left|\bigcup_{w>v} S_w\right| = n\sum_{w>v} p_w \leq Cn\sqrt{\ell_n}p_{\ell_n}. \tag{3.6}$$

Note also that $p^2_{\ell_n} \leq n^{-2+\delta}$ for any $\delta > 0$ and n sufficiently large. Taking expectations in 3.5, applying a union bound over the $2(\log n)^{3/2}$ values $v \in [\ell_n, r_n]$,

$$\mathbb{P}(S_v \neq \widetilde{S}_v \text{ for some } v \geq \ell_n) \leq Cn(\log n)^{5/2}p^2_{\ell_n} \to 0. \tag{3.7}$$

□

Putting everything together:

Theorem 3.2 (1.1)**.** *$d_{TV}(V_n, \widetilde{V_n}) \leq n^{-1+\delta}$ for any $\delta > 0$ as $n \to \infty$.*

Proof. Proposition 3.1 shows that V_n and $\widetilde{V_n}$ can be coupled to agree with high probability for all values v larger than ℓ_n, while Lemma 2.1 shows that they also agree with high probability (even without coupling) for all values $v \leq \ell_n$ (since they always contain the latter values with high probability). □