What is a tight lower bound on the coupon collector time?

In the classic Coupon Collector's problem, it is well known that the time T necessary to complete a set of n randomly-picked coupons satisfies $E[T] \sim n \ln n, Var(T) \sim n^2$, and $\Pr(T > n \ln n + cn) < e^{-c}$.

This upper bound is better than the one given by the Chebyshev inequality, which would be roughly $1/c^2$.

My question is: is there a corresponding better-than-Chebyshev *lower bound* for T? (e.g., something like $\Pr(T < n \ln n - cn) < e^{-c}$) ?

probability | probability-inequalities | coupon-collector-problem

edited Feb 27 '16 at 20:19 whuber ♦ 189k 30 401 745

asked Mar 2 '11 at 3:58 David 168 8

An obvious lower bound is $\Pr(T < n) = 0$, but I guess you're aware of that... – onestop Mar 2 '11 at 8:56

4 Answers

I'm providing this as a second answer since the analysis is completely elementary and provides exactly the desired result.

Proposition For $c > 0$ and $n \geq 1$,

$$\mathbb{P}(T < n \log n - cn) < e^{-c} .$$

The idea behind the proof is simple:

1. Represent the time until all coupons are collected as $T = \sum_{i=1}^{n} T_i$, where T_i is the time that the ith (heretofore) ***unique*** coupon is collected. The T_i are geometric random variables with mean times of $\frac{n}{n-i+1}$.
2. Apply a version of the Chernoff bound and simplify.

Proof

For any t and *any* $s > 0$, we have that

$$\mathbb{P}(T < t) = \mathbb{P}(e^{-sT} > e^{-st}) \leq e^{st}\mathbb{E}e^{-sT} .$$

Since $T = \sum_i T_i$ and the T_i are independent, we can write

$$\mathbb{E}e^{-sT} = \prod_{i=1}^{n} \mathbb{E}e^{-sT_i}$$

Now since T_i is geometric, let's say with probability of success p_i, then a simple calculation shows

$$\mathbb{E}e^{-sT_i} = \frac{p_i}{e^s - 1 + p_i}.$$

The p_i for our problem are $p_1 = 1, p_2 = 1 - 1/n, p_3 = 1 - 2/n$, etc. Hence,

$$\prod_{i=1}^{n} \mathbb{E}e^{-sT_i} = \prod_{i=1}^{n} \frac{i/n}{e^s - 1 + i/n}.$$

Let's *choose* $s = 1/n$ and $t = n \log n - cn$ for some $c > 0$. Then

$$e^{st} = ne^{-c}$$

and $e^s = e^{1/n} \geq 1 + 1/n$, yielding

$$\prod_{i=1}^{n} \frac{i/n}{e^s - 1 + i/n} \leq \prod_{i=1}^{n} \frac{i}{i+1} = \frac{1}{n+1} .$$

Putting this together, we get that

$$P(T < n\log n - cn) \leq \frac{n}{n+1} e^{-c} < e^{-c}$$

as desired.

edited Mar 5 '11 at 14:34

answered Mar 5 '11 at 14:28

cardinal ♦

20.5k 7 77 112

That's very nice and just what the doctor ordered. Thank you. – David Mar 6 '11 at 5:00

@David, just curious: What is the intended application? – cardinal ♦ Mar 6 '11 at 5:01

Long story. I'm trying to prove a lower bound for the mixing time of a Markov chain that I've cooked up in order to analyze the running time of an algorithm I'm interested in - which turns out to reduce to lower-bounding the c.collector problem. BTW, I had been playing around with trying to find exactly this kind of Chernoff-style bound, but hadn't figured out how to get rid of that product in i. Good call choosing $s = 1/n$:-). – David Mar 6 '11 at 5:23

@David, $s = 1/n$, while almost certainly suboptimal, seemed like the obvious thing to try since that gave $e^{st} = ne^{-c}$, which is the same term as the one obtained in the derivation for the upper bound. – cardinal ♦ Mar 6 '11 at 17:48

Request: The proof I've given above is my own. I worked on it out of enjoyment, since the problem intrigued me. However, I make no claim to novelty. Indeed, I cannot imagine that a similar proof using a similar technique doesn't already exist in the literature. If anyone knows of a reference, **please** post it as a comment here. I would be **very** interested to know of one. – cardinal ♦ Mar 7 '11 at 4:26

Although @cardinal has already given an answer that gives precisely the bound I was looking for, I have found a similar Chernoff-style argument that can give a stronger bound:

Proposition:

$$Pr(T \leq n\log n - cn) \leq \exp(-\frac{3c^2}{\pi^2}) .$$

(this is stronger for $c > \frac{\pi^2}{3}$)

Proof:

As in @cardinal's answer, we can use the fact that T is a sum of independent geometric random variables T_i with success probabilities $p_i = 1 - i/n$. It follows that $E[T_i] = 1/p_i$ and $E[T] = \sum_{i=1}^{n} E[T_i] = n\sum_{i=1}^{n} \frac{1}{i} \geq n\log n$.

Define now new variables $S_i := T_i - E[T_i]$, and $S := \sum_i S_i$. We can then write

$$\Pr(T \leq n\log n - cn) \leq \Pr(T \leq E[T] - cn) = \Pr(S \leq -cn)$$

$$= \Pr\left(\exp(-sS) \geq \exp(scn)\right) \leq e^{-scn} E\left[e^{-sS}\right]$$

Computing the averages, we have

$$E[e^{-sS}] = \prod_i E[e^{-sS_i}] = \prod_i \frac{e^{s/p_i}}{1 + \frac{1}{p_i}(e^s - 1)} \leq e^{\frac{1}{2}s^2 \sum_i p_i^{-2}}$$

where the inequality follows from the facts that $e^s - 1 \geq s$ and also $\frac{e^z}{1+z} \leq e^{\frac{1}{2}z^2}$ for $z \geq 0$.

Thus, since $\sum_i p_i^{-2} = n^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \leq n^2\pi^2/6$, we can write

$$\Pr(T \leq n\log n - cn) \leq e^{\frac{1}{12}(n\pi s)^2 - scn}.$$

Minimizing over $s > 0$, we finally obtain

$$\Pr(T \leq n\log n - cn) \leq e^{-\frac{3c^2}{\pi^2}}$$

edited Mar 8 '11 at 1:58

cardinal ♦

20.5k 7 77 112

answered Mar 7 '11 at 17:56

David

168 8

(+1) Modulo a couple of minor typos, this is nice. Expanding around something close to the mean as you've done often works better. I'm not surprised to see the higher order convergence in light of the asymptotic results. Now, if you show a similar such upper bound, that proves $(T - n\log n)/n$ is *subexponential* in the terminology of Vershynin, which has many implications regarding measure concentration. – cardinal ♦ Mar 7 '11 at 18:47

The argument doesn't seem to generalize directly to the upper bound. Exchanging c for $-c$ (and s for $-s$), one can follow the same steps up to the point of calculating $E[e^{sS}] \leq \prod_i \frac{e^{-s/p_i}}{1-\frac{s}{p_i}}$. At this point, however, the best I can do is to use

$\frac{e^{-z}}{1-z} \leq \exp(\frac{z^2}{2(1-z)})$, which still leaves

$$E[e^{sS}] \leq e^{\frac{1}{2}s^2 \sum_i \frac{p_i^2}{(1-s/p_i)}}$$

and I don't know what to do with this – David Mar 8 '11 at 3:27

1 Interestingly enough, though, the entire argument (for the lower bound) seems to work not only for the coupon collector problem, but for *any* sum of non-identical, independent geometric variables with bounded variance. Specifically: given $T = \sum_i T_i$, where each T_i is an independent GV with success probability p_i, and where $\sum_i p_i^{-2} \leq A < \infty$, then

$$\Pr(T \leq E[T] - a) \leq e^{-\frac{a^2}{2A}}$$

– David Mar 8 '11 at 3:35

Important Note: I've decided to remove the proof I gave originally in this answer. It was longer, more computational, used bigger hammers, and proved a weaker result as compared to the other proof I've given. All around, an inferior approach (in my view). If you're *really* interested, I suppose you can look at the edits.

The asymptotic results that I originally quoted and which are still found below in this answer do show that as $n \to \infty$ we can do a bit better than the bound proved in the other answer, which holds for *all* n.

The following *asymptotic* results hold

$$\mathbb{P}(T > n \log n + cn) \to 1 - e^{-e^{-c}}$$

and

$$\mathbb{P}(T \leq n \log n - cn) \to e^{-e^{c}} \ .$$

The constant $c \in \mathbb{R}$ and the limits are taken as $n \to \infty$. Note that, though they're separated into two results, they're pretty much the same result since c is not constrained to be nonnegative in either case.

See, e.g., Motwani and Raghavan, *Randomized Algorithms*, pp. 60--63 for a proof.

Also: David kindly provides a proof for his stated upper bound in the comments to this answer.

edited Mar 7 '11 at 13:01

answered Mar 2 '11 at 13:50

cardinal ♦

20.5k 7 77 112

Yes, it holds for every fixed n. A (very simple) proof can be found, for instance in Levin, Peres and Wilmer's book Markov Chains and Mixing Times, Proposition 2.4. The proof doesn't work for the lower bound, though. – David Mar 3 '11 at 4:41

1 In fact, I might as well transcribe the proof here: "Let A_i be the event that the i-th [coupon] type does not appear among the first $n \log n + cn$ coupons drawn. Observe first that $P(\tau > n \log n + cn) = P(\cup_i A_i) \leq \sum_i P(A_i)$. Since each trial has probability $1 - n^{-1}$ of not drawing coupon i and the trials are independent, the right-hand side above is bounded above by $\sum_i (1 - 1/n)^{n \log n + cn} \leq n \exp(\frac{n \log n + cn}{n}) = e^{-c}$, proving (2.7)." – David Mar 3 '11 at 5:17

@David, that's nice and simple enough. I quickly played with expanding the inclusion-exclusion formula out by another term, but didn't get anywhere quickly and haven't had time to look at it further. The event $\{T < t_n\}$ is equivalent to the event that no coupons are left after t_n trials. There should be a martingale associated with that. Did you try Hoeffding's inequality on the (presumed) associated martingale? The asymptotic result suggest strong measure concentration. – cardinal ♦ Mar 3 '11 at 18:48

@David, there's a sign flip in your proof above, but I'm sure that's obvious to other readers too. – cardinal ♦ Mar 3 '11 at 18:50

@David, please see my other posted answer to your question. The method is different than the upper bound you give, but the tools employed are nearly as elementary, in contrast to the answer I gave here. – cardinal ♦ Mar 5 '11 at 14:40

Benjamin Doerr gives (in the chapter " Analyzing Randomized Search Heuristics: Tools from Probability Theory" in the book "Theory of Randomized Search Heuristics", see the link for an online PDF) a somewhat simple proof of

Proposition Let T be the stopping time of the coupon collection process. Then $\Pr[T \leq (1 - \epsilon)(n-1) \ln n] \leq e^{-n^{\epsilon}}$.

This seems to give the desired asymptotics (from @cardinal's second answer), but with the advantage of being true for all n and ϵ.

Here is a proof sketch.

Proof Sketch: Let X_i be the event that the i-th coupon is collected in the first t draws. Thus, $\Pr[X_i = 1] = (1 - 1/n)^t$. The key fact is that the X_i are negatively correlated, for any $I \subseteq [n]$, $\Pr[\forall i \in I, X_i = 1] \leq \prod_{i \in I} \Pr[X_i = 1]$. Intuitively, this is fairly clear, as knowing that the i-th coupon in the first t draws would make it less likely that the j-th coupon is also drawn in the first t draws.

One can prove the claim, but enlarging the set I by 1 at each step. Then it reduces to showing that $\Pr[\forall i \in I, X_i = 1 | X_j = 1] \leq \Pr[\forall i \in I, X_i = 1]$, for $j \notin I$. Equivalently, by averaging, it reduces to showing that $\Pr[\forall i \in I, X_i = 1 | X_j = 0] \geq \Pr[\forall i \in I, X_i = 1]$. Doerr only gives an intuitive argument for this. One avenue to a proof is as follows. One can observe that conditioned on the j coupon coming after all of the coupons in I, that the probability of drawing a new coupon from I after

drawing k so far is now $\frac{|I|-k}{n-1}$, instead of the previous $\frac{|I|-k}{n}$. So decomposing the time to collect all coupons as a sum of geometric random variables, we can see that conditioning on the j-coupon coming after I increases the success probabilities, and thus doing the conditioning only makes it more likely to collect the coupons earlier (by stochastic dominance: each geometric random variable is increased, in terms of stochastic dominance, by the conditioning, and this dominance can then be applied to the sum).

Given this negative correlation, it follows that $\Pr[T \leq (1-\epsilon)(n-1)\ln n] \leq (1-(1-1/n)^t)^n$, which gives the desired bound with $t = (1-\epsilon)(n-1)\ln n$.

edited Jul 16 '13 at 8:51

answered Jul 16 '13 at 8:36

miforbes
121 3